The Missing Link: Estimating the Impact of Incentives on Teacher Effort and Instructional Effectiveness Using Teacher Accountability Legislation Data

Tom Ahn
University of Kentucky

Teacher effort, a critical component of education production, has been understudied in the literature because of measurement difficulties. I use a principal-agent model, North Carolina data, and the state's accountability system that awards cash for school-level academic growth to distill effort from teacher absence and capture its effect. I find low effort at low and high probabilities of bonus receipt, high effort when the bonus outcome is in doubt, and free-ridership. Teachers respond to incentives, and effort affects achievement. Policy simulations with individual-level incentives eliminate free-rider effects but reduce effort by pushing teachers into the tails of the probability of bonus receipt distribution.

I. Introduction

Extensive empirical evidence exists to suggest that teacher quality is one of the most important determinants of a student's success in his or her academic career (see, e.g, Sanders and Rivers 1996; Darling-Hammond 2000; Rivkin, Hanushek, and Kain 2005). Observable teacher characteristics such as experience, advanced degrees, and credentials have been identified as proxies for teacher quality. Previous studies have found these variables to have weak to moderate positive correlation with higher student achievement. While these characteristics explain part of the impact of teachers, they fail to account for all of the observed variation in achievement. Researchers have often modeled this residual unaccounted-for effect of teachers as individual-specific constants in fixed-effects models (see, e.g., Rockoff 2004; Clotfelter, Ladd, and Vigdor 2007b; Goldhaber and Anthony 2007).

In the meantime, policy makers have used accountability systems to introduce "market forces" to improve standardized test scores. If sanctions or bonuses are expected to raise student achievement, these punishments or rewards must be designed to affect some portion of teacher quality that is not easily observable in an administrative data set. This paper uncovers

[*Journal of Human Capital,* 2013, vol. 7, no. 3]

this important and unexplored dimension of teacher quality as effort, the one endogenous characteristic common to all teachers that can be immediately influenced by incentives.

While there is considerable debate on whether accountability systems help students as intended, there is evidence that teachers and administrators respond to incentives. Studies have shown that performance incentives for teachers can lead to higher academic performance, states with stricter accountability standards have been associated with higher test scores, and high-stakes testing improves student achievement compared to low-stakes testing. The accumulated evidence shows that incentives can raise test scores.[1] This study adds to the literature documenting the effectiveness of accountability systems by estimating a structural model of education production with a pay for performance teacher compensation component.

Accountability systems may raise test scores, but test score increases by themselves do not always reflect actual education production. Test scores can be raised in one of three ways. The first method is to alter the way in which teachers are hired or fired. When "better" teachers are selected or attracted to start and teachers who prove to be ineffective are dismissed, students will receive better instruction, leading to higher test scores (see, e.g., Rockoff et al. 2008; Hanushek 2009). While changing the teacher labor market in a fundamental way will most likely have the largest impact, whether accountability systems can effect this type of change remains an open question.[2]

The second way is to change teacher or school behavior unrelated to measurable effort exertion. Some of these changes lead to no actual education production. Examples of such behaviors include classifying marginal students as disabled, suspending them to prevent testing, or altering students' answer sheets (see Cullen and Reback 2002; Figlio and Getzler 2002; Jacob and Levitt 2003; Jacob 2005; Figlio 2006). Some behavior may or may not lead to education production. "Teaching to the test," for example, may take valuable instruction time away from education production if the standardized test is poorly designed, or it may help teachers to create a good curriculum and impart necessary knowledge to students if the test is well designed (see Grissmer and Flanagan 1998; Hanushek and Raymond 2005). Some of these behaviors will systematically bias the relationship between the outcome and the inputs, making raw exam scores an imprecise measure of student achievement.[3]

[1] See Clotfelter and Ladd (1996), Carnoy and Loeb (2002), Figlio and Kenny (2007), Jacob (2007), Vigdor (2008), Lavy (2009), and Dee and Jacob (2011). Some recent experimental studies have found no impact of incentives. See Sec. II.

[2] See the discussion in Sec. VI.B.

[3] It should be noted that this bias is in addition to the already noisy test score as a measure of student achievement. If combined with a risk-averse teacher, the inherent noise may disincentivize teachers from exerting effort, especially at smaller schools, where this noise component would be larger. See Kane and Staiger (2002).

Finally, the third way is to increase teacher effort. An increase in effort will lead to real education production. Many studies have discussed the importance of teacher effort, and some have made the connection between teacher absence and teacher effort or student achievement. Studies generally find negative correlation between an increase in "market forces" (such as decreased job security) and absenteeism. Studies that explicitly look at absence and academic achievement find a small to moderate decrease in test scores with increasing teacher absence (see, e.g., Duflo and Hanna 2006; Bradley, Green, and Leeves 2007; Clotfelter, Ladd, and Vigdor 2007a; Miller, Murnane, and Willett 2008; Jacob 2010). While using absence as a direct proxy for effort yields some suggestive results, I propose a method of isolating the true effort response of teachers to incentives using a unique data set collected in North Carolina that tracks the academic history, demographic information, and teacher and peer exposure of students in the public school system. This data set, along with the unique teacher incentive system that pays out bonuses for school-level year-over-year improvement in standardized test scores, makes it possible to distill the teacher's level of effort, identify her effort response to incentives, gauge the impact of effort on achievement, and evaluate the efficacy of the accountability policy.[4]

Recent research has focused on the distributional changes in achievement (see Booher-Jennings 2005; White and Rosenbaum 2007; Neal and Schanzenbach 2010). In addition to the change in distribution of resources or effort, which creates winners and losers among the student population, there may still be net positive or negative effects, depending on aggregate effort level change due to the incentive policy. This study will examine efficiency as well as distributional implications of accountability policies.

The key insight from the theoretical model and econometric analysis is that the conventional wisdom of eliminating free-ridership to increase the effectiveness of incentives may actually be counterproductive when the accountability system uses strict thresholds (based on year-over-year test score gains) to make binary decisions on whether teachers are rewarded as a group.[5]

Policy simulations showcase the trade-off between the free-rider effect and the incentive effect generated by the accountability policy. Results shows that classroom-level incentives lower education achievement compared to school-level incentives. While negative free-rider effects are eliminated with individual-level incentives, isolating teachers has the unde-

[4] Another method is to induce more student effort. Evidence of efficacy remains mixed (see Angrist and Lavy 2002).

[5] This issue can be understood as a principal-agent problem, with the state government as the principal, teachers as agents, and the accountability incentives as the all-or-nothing bonus contract. In the context of this study, school achievement is the cooperative team project that the bonus contract is based on, student- or classroom-level achievement is the noisy individual output, and teacher absence is the imperfect signal of worker effort.

sirable effect of pushing many teachers toward the tails of the probability distribution, where incentive effects of the bonus are minimized.

Section II details the North Carolina accountability system. Section III presents a simple theoretical model. Section IV describes the data, and Section V introduces the econometric model. I present the results in Section VI. Section VII discusses the implications of altering the bonus system. I present conclusions in Section VIII. Proofs for the theoretical model are presented in Appendix A. Robustness checks and minor regression results are presented in online Appendix B.

II. The Accountability System

A. *The North Carolina ABC System*

The North Carolina accountability program (also known as ABC) began in the 1995/96 academic year.[6] While the system has grown in complexity and has gone through minor alterations in the details of execution, the principal mechanism of offering cash incentives for student achievement gains has remained unchanged for more than a decade.[7] In contrast to a level model such as No Child Left Behind (NCLB) that rewards absolute achievement, the focus of the North Carolina system has always been on growth of scores from the previous year.

North Carolina public school students in grades 3–8 must take end-of-grade (EOG) exams in reading and mathematics. The test is on a developmental scale, allowing comparison of scores across grades. Using the formula defined below, the North Carolina Department of Public Instruction (NCDPI) determines the required achievement gains for each school based on the school's students' performance last year on the EOG exams. That is, the difference in test scores is initially captured at the student level, allowing transfer students to be included in the school achievement calculations. The system is two-tiered, with teachers in schools making "expected growth" receiving a \$750 bonus and teachers in schools making "high growth" earning \$1,500. "Expected" and "high" growth are defined below.

The achievement gain threshold used in the calculation of bonus eligibility is defined as

$$\Delta y_{mgst} = \Delta y_{gs94} + b_1 \mathrm{ITP}_{gst} + b_2 \mathrm{IRM}_{mgst}.$$

The Δy_{mgst} term is the required change in the test score in subject m for students in grade g in year t in school s compared to the score last year,

[6] The acronym stands for strong Accountability, teaching the Basics, and emphasis on local Control.

[7] For instance, middle school and high school achievement gains were measured starting in 1997/98. For a complete description of the incentive system as well as the high school criterion calculation, see Vigdor (2008).

and Δy_{gs94} is the average change in standardized test scores for all North Carolina students in 1993/94 compared to the scores from 1992/93.

The second and third terms on the right-hand side are "correction factors." The ITP_{gst} term is the index of true proficiency, and the IRM_{mgst} term is the index for the regression to the mean. The two terms are meant to adjust test score goals for shocks in performance of the school's students last year.[8]

With this formula, for a school with G tested grades, $2G$ thresholds are produced each year (math and reading), which are compared to the actual average test score improvement at the school. The school scores and threshold scores are differenced, standardized, weighted by the number of students in each grade, and summed across grades and subjects. That is, each school receives a z-score centered at zero and a standard deviation of one. If this z-score, termed "expected growth composite," is greater than zero, teachers in the school receive a $750 bonus. Teachers in schools that make expected growth yet fail to test more than 98 percent of eligible students are exempted from the bonus. The procedure is repeated after increasing the growth threshold by 10 percent to generate the "high-growth composite." Teachers in schools that make high growth receive an additional cash bonus of $750. Therefore, teachers in a school with exceptional test score growth scores can earn as much as $1,500.

B. *Other Teacher Incentive Systems*

There is no clear consensus on the effectiveness of pay for performance systems in the literature. Some studies find that bonus incentives can have large positive impacts, while others find little to no impact, even with substantial bonus payouts.

In the United States, teacher or school incentive programs of various types exist (or had existed before expiring) in 40 states.[9] Large incentive programs or experiments that evaluated pay for performance were imple-

[8] ITP is obtained by subtracting the sum of the 1994/95 state average reading and math test scores from the sum of the average reading and math scores of students in grade $g-1$, school s, in year -1. Therefore, one ITP value is generated per school. The b_1 term varies by grade and subject but is positive in all cases. Thus, schools that performed better compared to the state standard last year must achieve higher growth this year to qualify for the cash bonus, all else equal. IRM is generated by taking the difference of the average test scores of students in grade $g-1$, school s, in year $t-1$ and the 1994/95 state average test scores for each subject. Therefore, two IRM values are generated per school. The b_2 term varies by grade and subject but is negative in all cases. IRM is meant to account for the possibility that some students may have performed particularly well (or poorly) as a statistical anomaly on a particular test and will return to expected performance the next year. Thus, if some students performed abnormally well on the math test in year $t-1$, the state does not place overly harsh expectations on the school to replicate this feat in year t. I remain agnostic about the effectiveness of ITP and IRM in accounting for these macro-level shocks, but they may help to correct for some measurement error, especially in small schools.

[9] Some states have had several different incentive programs in place at different times. For example, Florida has had at least three different systems (E-Comp, Special Teachers Are Rewarded, and Merit Award Program [MAP]). Recently, in 2011, a new pay for performance law was introduced in Florida.

mented in Colorado (Denver), Florida, Illinois (Chicago), Minnesota, New York City, North Carolina, Tennessee, and Texas (Dallas and Houston).[10]

The Denver ProComp system awarded up to 6.4 percent of the "index salary" (in 2010, $37,551) to teachers on the basis of school-level proficiency or school-level growth in test scores. Teachers could earn bonuses in a variety of other ways as well, including attaining advanced degrees or certification and working in hard-to-staff schools. Goldhaber and Walch (2012) found that students taught by both ProComp participant and non-participant teachers made significant gains. In Florida, the most recently administered incentive program (MAP) awarded bonuses that ranged from 5 percent to 10 percent of a district teacher's average annual salary. Each district had discretion over the proportion of teachers who received the bonus, and at least 60 percent of the criterion had to be based on student proficiency or gains. A Chicago experiment that tested for loss aversion by teachers (of having to "give back" the bonus for poor student performance, compared to the traditional bonus) in a pay for percentile framework showed a positive and significant impact of the loss-focused incentives (see Jacob 2005). In Minnesota, the Q-Comp system allowed schools to choose either school-level or individual-level incentives. A teacher could be awarded a maximum of $2,000 on the basis of several factors, such as peer observations, professional development credits, and student achievement. Sojourner, West, and Mykerezi (2011) found a small but positive and significant increase in student achievement. In New York City, a school-level randomized teacher incentive experiment that paid out approximately $3,000 per union teacher was conducted. Fryer (2011) found the bonus to have little impact on test scores. A large-scale incentive experiment in Nashville that ran from 2006 to 2009 grouped middle school teachers into teams and evaluated them on the basis of value-added measures of student performance. Each teacher could earn approximately $6,000, but Springer et al. (2012) found no significant effect on test scores or changes to teaching practices.[11] Ladd (1999) analyzed a Dallas merit pay program (1991–95) on the basis of schoolwide performance that awarded $1,000 annual bonuses and found that the bonus incentives resulted in test score gains. Imberman and Lovenheim (2012) investigated a Houston rank-order tournament incentive pay program (2006–10) that paid out substantial bonuses (up to $7,000) to teachers grouped at the grade-subject level. The authors found that achievement increased, with smaller groupings making larger gains, pointing to possible free-rider effects.

Internationally, well-known studies of incentive programs were conducted in India, Israel, and Kenya. Many of these programs were short-

[10] This is by no means an exhaustive list, but it does represent a cross section of different types of incentive programs.

[11] In the New York City and Nashville systems, it appears that a majority of teachers did not understand how the bonus outcome was determined, which could have contributed to the lack of observed impact.

term (or one-shot) evaluations that showed that teacher incentives can raise student achievement.

Muralidharan and Sundararaman (2011) evaluated the impact of incentives on teachers in a large-scale randomized evaluation in a rural province in India. Bonuses were determined by percentage improvement in math and language test scores from the previous year, and average bonus awards amounted to about 3 percent of annual salaries. The study found that individual-level incentives were more effective in raising test scores compared to school-level incentives. In Israel, a schoolwide incentive program instituted in 1995 was studied by Lavy (2002). Bonuses were paid out on the basis of dropout rates, the average number of credits taken per student, and the fraction of students receiving matriculation certificates. The top-third performing schools were awarded the incentive, and the school divided this between salary bonuses for teachers and school improvement uses. Teacher bonus amounts ranged from $250 to $1,000. The study found that these school-based incentives raised test scores, decreased dropout, and increased the fraction of students receiving matriculation certificates. Lavy (2009) studied another Israeli program in 2001 that paid bonuses to teachers of grades 7–12 who taught English, Hebrew and Arabic, math, and "other subjects" in a rank-order tournament. Bonus awards were paid out at the individual level and were large, ranging from $1,750 to $7,500. Teachers could win multiple awards per year. The incentives were shown to be effective in increasing test scores. Glewwe, Ilias, and Kremer (2010) studied school-level incentives in Kenya that paid awards amounting to approximately 20–40 percent of 1-month salaries to teachers on the basis of test score improvements in all subjects in the school. The authors found that test scores did increase, but they speculate that this was not due to an increase in teacher effort (as teacher attendance did not change) but to an increase in the number of test preparation sessions.

As we can see from the above examples, pay for performance can mean many different things. Measures of student outcomes can include test score growth (value added), proficiency level, school attendance, test participation, and graduation rate. Certain subjects, grades, students, and schools may be exempt from testing. Performance may be measured at the class, grade, subject, or school level. There may or may not be competition against other teachers or schools. Performance may be assessed alongside traditional "input" measures such as advanced degrees and certification. Bonuses can be all or nothing or pay for percentile. And of course, the amount of payouts can vary substantially among programs. With such differences in implementation, it is not surprising that different studies have reached different conclusions.

The North Carolina program can be characterized as narrowly focused on year-over-year test score improvements with modest bonus payouts. Teachers and schools were not in competition against each other, and the structure of the incentives was relatively simple and easy to understand.

Elementary schools were assessed on the growth of reading and mathematics test scores only, with schools rewarded if they managed to cross a strict threshold that was based on average growth rate in the base year across the entire state. The $1,500 maximum bonus was lower than average payouts in many other states. The program was relatively long-lived (10+ years) and was implemented statewide.

The simplicity of the program and the stability offered by its longevity and geographic breadth allow tractable theoretical and structural econometric models to be constructed that capture most of the salient features of the incentive program. The modest payouts allay some concerns about other possible influences that may drive results, such as large-scale coordination among teachers and schools across years or direct intervention by districts. The elements that make the North Carolina system easier to analyze are also what distinguishes the ABC system from other state systems. Implementation details matter, and caution should be exercised when comparing relative effectiveness among incentive programs.

III. Theoretical Model

Teacher j's utility function is defined as the difference between expected bonus and effort cost. I assume that teachers are risk neutral in bonus receipt. Teachers are differentiated by ability, $x_j \in [\underline{x}, \bar{x}]$, and effort, $e_j \in [\underline{e}, \bar{e}]$.[12] Effort e_j in the context of the model is not the absolute amount of effort that can be exerted by a teacher. Instead, $\underline{e}$ is the amount of effort a teacher will exert in the complete absence of incentive pressure, and $\bar{e}$ is the maximum effort a teacher will exert at "optimal" incentive pressure. The term B is the bonus that is paid to all teachers at the school upon qualifying under the state criterion, Cr. Because the bonus is determined at the school level, all teachers' efforts and abilities contribute to the probability of bonus receipt, and the incentive system may suffer from a free-rider problem:[13]

$$U_j = B \cdot \Pr(e_1, e_2, \ldots, e_J, x_1, x_2, \ldots, x_J, Cr) - C(e_j). \tag{1}$$

Define school average achievement in year t as Y_t:

$$Y_t = \frac{\sum_1^J y(x_{jt}, e_{jt}) \cdot w_{jt}}{J_t} + \epsilon_t, \tag{2}$$

where $y(x_{jt}, e_{jt})$ is the average classroom-level achievement in class/teacher j in year t, w_{jt} is class size weight (the ratio of the size of teacher j's class over

[12] I abstract away from student ability here, but the econometric model controls for observable student and peer characteristics.

[13] Teachers are assumed to be risk neutral with respect to the bonus receipt. If teachers are risk averse, the increase in uncertainty of outcome may dilute the effectiveness of the bonus scheme. On the other hand, risk aversion may be one argument put forth in favor of school-level incentives.

the average class size in the school), J_t represents the number of teachers at the school, and ϵ_t is a stochastic error term that captures school-level shocks such as inclement weather or flu outbreaks. If class sizes across the school are identical, $w_{jt} = 1$ for all j, and the ratio drops out. With $\overline{Y}_{t-1}$ defined as last-year schoolwide achievement, the probability of qualifying for the bonus is defined by

$$\Pr_t = F((Y_t - \overline{Y}_{t-1}) - Cr). \tag{3}$$

I make the following assumptions.

Assumption 1. The function $F(\cdot) \in [0, 1]$ is twice differentiable and $F'(\cdot) \geq 0$.

Assumption 2. The effort cost function $C(e)$ is twice differentiable, with $C'(\cdot) > 0$ and $C''(\cdot) \geq 0$.

Assumption 3.

$$\frac{Bw_{jt}}{J_t}[F''(\cdot)a'(\cdot) + F'(\cdot)a''(\cdot)] \leq C''(\cdot).$$

Assumption 4. Defining Y_{-j} as the school average achievement without teacher j's class, for some $J > J^*$, there exists a high value of Y, Y^H, and some low value, Y^L, such that $\Pr_t | Y_{-j} > Y^H \rightarrow 1$ and $\Pr_t | Y_{-j} < Y^L \rightarrow 0$ for the entire range of e_j and x_j.

Proposition 1. Given equations (1)–(3), assumptions 1–4, and $\{x_1, \ldots, x_J, w_1, \ldots, w_J, B, Cr, J\}$, there exists an interior pure-strategy Nash equilibrium in effort, $\{e_1, e_2, \ldots, e_J\}$.

Assumption 3 ensures global concavity of expected utility. Assumption 4 is a statement about the limitation of effort within a large school. One teacher cannot unilaterally determine the bonus receipt of the entire school. If all other teachers shirk (and/or are low-ability), the best response of teacher j is also to shirk. On the other hand, if all teachers are giving maximal effort (and/or are high-ability) such that the bonus is assured, the best response of teacher j is again to shirk. It is exactly when all other teachers are putting forth effort and the bonus outcome is in doubt that teacher j is also induced to give some positive amount of effort. I show that there can exist a free-rider problem in the incentive system in the following proposition.

Proposition 2. Assuming identical teachers and $Y^L < Y < Y^H$, a free-rider problem may exist.

As the number of teachers increases from J to $J + 1$, a teacher's effort is distributed over a larger population. Since her payoff remains constant, she will be induced to lower her level of effort. In this sense, free-rider problems always exist, unless incentives are reduced to individual-level bonuses.

Imposing identical teachers in proposition 2 is a strong assumption because it ignores the second force in the model that affects teacher effort: the distribution of teacher ability across the school. An increase in the number of teachers in the model necessarily implies an additional class.[14] If the new teacher's ability, x_{J+1}, is significantly different from the school average teacher ability (and if the school is small enough), this can change the probability of bonus receipt for the school, which can have a large impact on effort exertion across the school.

For the empirical model, I assume that the classroom average achievement is generated as

$$y(x_{jt}, e_{jt}) = \exp(x_{jt})e_{jt}$$

and that the effort cost function is exponential.[15] The first-order condition from the teacher utility function can be written as

$$J^{-1} \cdot B \cdot F'(Y) \cdot \exp(x_j) \cdot w_j = C'(e_j).$$

I assume that the probability distribution is standard normal. When natural logs are taken, the first-order condition becomes

$$e_j^* = \gamma + x_j + \ln w_j - \ln J + \ln \phi(Y),$$

where γ is a normalizing term. In this way, the free-rider effect arising from an increase in J is separated from $\phi(\cdot)$, the probability density function (pdf) of the bonus receipt outcome. This will greatly simplify the econometric specification.[16]

Having found that school incentive policy may suffer from a free-rider problem, a simple solution would seem to be to go from a schoolwide incentive to a classroom-level (individual) incentive in which teachers are judged only on their students' performance. The first-order condition shows that the $-\ln J$ term would equal zero when $J = 1$, thus eliminating the free-rider effect altogether. However, I show below that moving to this noncooperative criterion will not necessarily increase effort exertion of teachers. In order to simplify the discussion, I add one more assumption.

Assumption 5. When $J = 1$, there exists some high value of x, x^H, and some low value, x^L, such that

$$\frac{\partial F(\cdot|x_j \geq x^H)}{\partial e} \cong 0, \quad \frac{\partial F(\cdot|x_j \leq x^L)}{\partial e} \cong 0.$$

14 Assuming that J or w_j is under the control of the administrator significantly complicates the model. Implications of selecting an "optimal" J or w_j are more fully explored at the end of this section.

15 Alternative forms of the effort cost function yielded no qualitative differences. The functional form of the classroom average achievement is admittedly ad hoc to allow for a linear regression (after taking natural logs) in the empirical section.

16 It should be noted that J still exists within $\phi(Y)$. Therefore, the ln J term does not isolate the full impact of J in the teacher's behavior, only the direct effect of diluting the effect of one teacher across a larger population. The J term embedded within y serves as a countervailing force by averaging across class scores to pull school average scores near the middle of the pdf (and thus inspiring higher effort exertion).

This assumption is a statement about the limitations of the incentives. For the case of a single-classroom school, a teacher's location on the bonus distribution is more strongly determined by ability than by effort induced by incentives. That is, it is not possible to change a low-ability teacher into a substantially more effective teacher simply by offering more money. Similarly, a high-ability teacher will not turn into a significantly less able teacher because of a smaller bonus (in the form of a reduced probability of bonus receipt). In the model, an isolated teacher with very low ability $(x_j < x^L)$ will have close to zero probability of qualifying for the bonus. Maximum effort induced by incentives cannot improve student achievement enough for this teacher to get beyond the flat part of the distribution on the left tail. Similarly, an isolated teacher with very high ability $(x_j > x^H)$ will have a bonus probability approaching one. A reduction in teacher effort due to decreased incentives will not lower students' test scores enough to move her off of the right tail of the bonus distribution.

Proposition 3. It is possible for effort to decline when the test score aggregation for incentives changes from school average to classroom average.

While decreasing J to one will increase effort by eliminating the free-rider problem, moving from schoolwide average to classroom average will move each teacher to a different point on $F(\cdot)$. Whereas a teacher may have been at a point on the distribution where her marginal effort can make a difference in the school average criterion, the classroom average criterion may place her at the tails of the distribution where additional effort exertion makes little difference in changing the probability of bonus receipt. This change in optimal effort exertion based on a teacher's placement on the distribution of the bonus receipt is critical in the analysis of any policy that attempts to increase student achievement by altering the incentive system.

This property is best demonstrated by looking at a school with two teachers. Assume that teacher 1 has $x_1 > x^H$ and teacher 2 has $x_2 < x^L$. These teachers' optimal solutions are demonstrated in figures 1 and 2. The thick lines represent indifference curves of teachers, and utility increases up and to the left. When the bonus is determined at the class level, the optimum solutions for both teachers are at the corner $(e_1 = e_2 = \underline{e})$ because neither teacher can significantly increase the probability of bonus receipt by exerting extra effort. If a bonus is awarded for joint performance, the optimum solutions move to interior points, and $e_i > \underline{e}$ for $i = 1, 2$, as marginal effort exertion for both teachers will now increase the probability of bonus receipt. I assume that the higher-ability teacher's marginal effort application is more effective compared to that of the lower-ability teacher, but the solution holds if I assume the opposite.

The results indicate that a school average incentive system with free-rider problems may be preferable to a classroom average bonus system depending on the distribution of teacher ability within the school. If schools are composed mostly of homogeneous teachers, the free-rider ef-

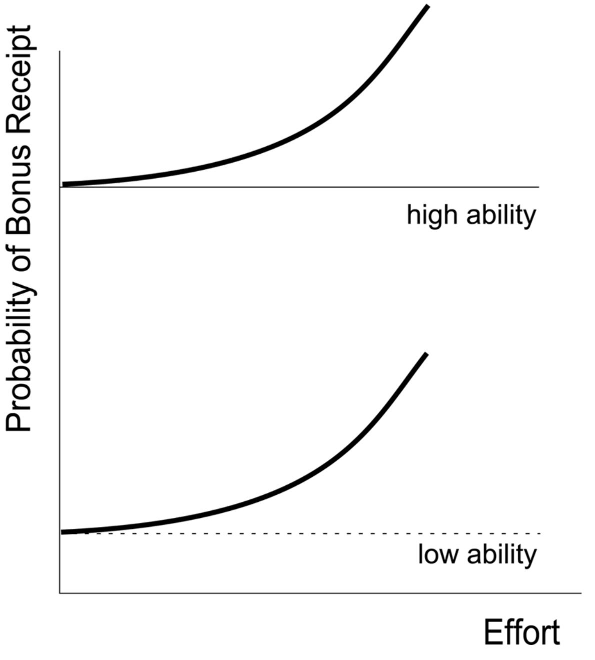

Figure 1.—Noncooperative regime

fect may dominate the incentive effect. If the bonus depends on classroom achievement and if the state criterion is set correctly such that teachers are moved to the "middle" of the probability distribution, they may be induced to exert more effort, and student achievement may increase. On the other hand, if the variance in teacher ability is high within the school, a school average incentive policy may actually induce greater effort and increased achievement because the free-rider effect is offset by the increase in incentive effects. This property does not rely on the traditional arguments against a competitive criterion: that it may deter teachers from sharing knowledge or working together. The only motivating factor is that bonus receipt is judged on joint achievement.[17]

The schoolwide criterion is a blunt instrument to gather teachers into a narrower band on the bonus receipt probability. When success depends on joint performance, teachers at the left and right extremes are grouped together and "gathered" to the middle of the distribution. Tightly grouping teachers will increase the power of the incentives only if the spot to which they are gathered allows them to positively affect the probability of bonus receipt through the application of effort. This potential positive

[17] If cooperation or a collaborative working environment is a positive nonmonetary job characteristic that exists (or is stronger) because of the schoolwide bonus, moving to a classroom standard would have an even greater (negative) impact. If cooperation is captured by absence (i.e., one of the ways in which teachers exert effort is to share didactic techniques), the interpretation from the model would remain the same.

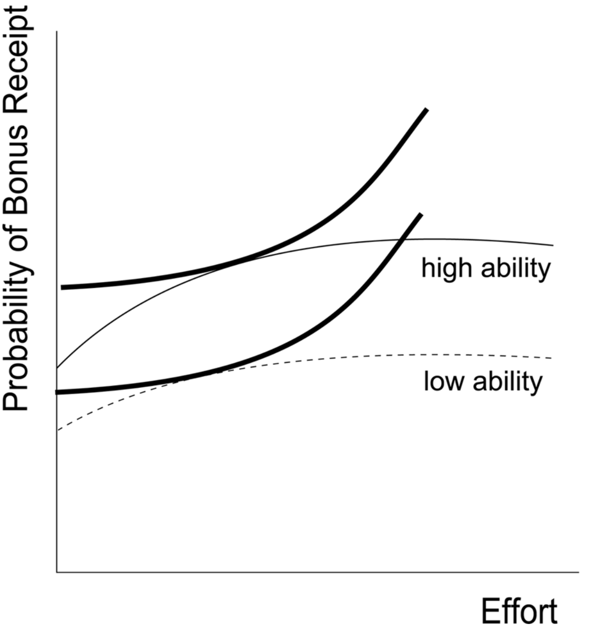

Figure 2.—Cooperative regime

effect of the joint performance criterion must be weighed against the free-rider effects inherent in the system. The policy simulation section explores these implications by changing the test score aggregation rule from school level to grade level to classroom level.

It is possible that the accountability system affects teacher behavior in other ways. For instance, teachers may attempt to coordinate effort such that they do not overimprove student achievement in year t, to the detriment of gains in year $t + 1$. If year-over-year gains matter and teachers have fine control over student achievement, teachers should invest just enough effort to receive the bonus each year and make "steady" progress by solving an intertemporal problem of effort exertion (see Macartney 2012). The likelihood of teachers playing a multiyear coordinated game may be of concern in smaller schools with low teacher mobility. In general, yearly turnover of teachers in North Carolina is almost 10 percent. While teachers may be engaging in the dynamic game, the bonus outcome across the 5-year data set was relatively volatile. If a school was categorized as making less than expected, expected, or high growth in the 1999/2000 academic year, the probability that the school would maintain its status across the 5-year period was approximately 8.5 percent. Across any 4-year span, the probability of a school maintaining its status was about 13 percent.

Other policy instruments.—In addition to the formal assumptions made above, I also do not allow the principal or the superintendent to alter class size by (1) adding more teachers (increasing J) holding student population constant or (2) having different class sizes for different teachers. While decreasing class size across the school or optimally distributing students to teachers will unequivocally increase academic achievement in a school with no accountability system in place, the incentive pressure introduced by the pay for performance system significantly complicates the impact of the above policy instruments.

The negative impact of increasing the number of teachers on effort suggests that the often-observed negative correlation between school size and academic achievement in the literature may be due in part to this free-rider impact. The existence of a free-rider effect implies that a given level of incentive pressure should be more keenly felt in smaller schools. This result means that accountability pressure should be geared toward affecting larger schools. Since a free-rider effect is stronger in larger schools, stronger pressure (or greater rewards/punishments) will be required to extract the same level of effort as in smaller schools. This result complicates attempts to increase academic achievement through class size reduction. Decreasing class size by adding more teachers may prove to be less effective than expected because the increase in the number of teachers exacerbates free-rider effects, pulling teacher effort and achievement down. This force will push against the increase in test scores from class size reduction.

Incentive pressure also plays a role in determining the ideal proportion of students to be distributed among teachers with varying ability. In general, the administrator will lean toward assigning more students to higher-ability teachers until average academic achievement across the school is maximized. At one extreme, when there is no pay for performance system in place, the principal may be induced to evenly distribute students across all teachers. At the other extreme, when incentive pressure is strong, the optimal solution may be to assign zero students to low-ability teachers. See Appendix A for a discussion of both cases.

IV. Data

I use an administrative data set for the North Carolina public school system from the academic years 1999/2000 to 2003/4.[18] The data set contains information on all public schools, students, teachers, and administrators in North Carolina. Since the data are collected annually and individuals can be matched across years, a relatively complete longitudinal picture of the entire public school system in North Carolina emerges, detailing

[18] The data, which are collected by the NCDPI, were made available by the North Carolina Education Research Data Center (http://www.pubpol.duke.edu/centers/child/nceddata center.html) at the Center for Child and Family Policy. While student- and teacher-level data are confidential, aggregate data and summary statistics are publicly available at the NCDPI website (http://www.ncpublicschools.org/reportstats.html). Post-2004 data have unreliable absence data, and pre-1999 data yield poorer teacher-student matches.

students' academic trajectories, peer interactions, and exposure to teachers.[19]

It should be noted that in the 2003/4 academic year, the North Carolina public school system was also subject to NCLB. I abstract away from impacts from NCLB sanctions. The criteria for making expected growth and adequate yearly progress (AYP) are largely uncorrelated. In fact, approximately 40 percent of schools in North Carolina pass under one regime yet fail under the other. In addition, two consecutive years of failures to make AYP are required before schools are subject to sanctions, and the sanctions that these schools would have been under—being forced to offer students the option to transfer to a different school—had very low uptake and did not affect teachers individually.[20]

There are two unique features of the data that I take advantage of to identify the effect of teacher effort. The first is that each student record is linked to a teacher. This permits the identification of a complete classroom, with information on the student, teacher, and peers, provided that student instruction is confined primarily to the self-contained classroom. While students in middle and high schools change teachers and peers each period, elementary school students are tied to a single classroom, where they are exposed to the same peers and teacher throughout the school day. Therefore, any effect of effort from the elementary school teacher should be isolated to her classroom. I focus on teachers of students in grades 3–5 in schools that have grade 5 as the highest grade. Each student in the data set has a pre- and postteacher exposure test score. Students in grades 4 and 5 use the previous year's test score as the preexposure score, and students in grade 3 take a separate exam at the start of the academic year. While teachers in charge of kindergarten through grade 2 contribute to students' academic growth, their efforts are not explicitly evaluated.

The data set tracks a student's performance year to year as long as he or she remains in the North Carolina public school system. Because of the need for at least 2 years of EOG exam results for each student to judge whether he or she has improved, students with only a single-year record are dropped.[21] Demographic characteristics such as sex, race, age, and parental education level are collected. I divide race into minority (black, Hispanic, American Indian) and white (and Asians). I divide parental edu-

[19] It should be noted that, on average, about 80 percent of all records of students are successfully matched with teachers. Charter schools were dropped because of inconsistent student-teacher match rates across the sample. In general, charter school students made up less than 3 percent of the state's student population during the sample years.

[20] While teachers may not be incentivized by NCLB pressure, it is possible that the principal may be affected. This may lead the school to change behavior, such as targeting certain teachers to certain classes. For more details, see Ahn and Vigdor (2013).

[21] These students may be transfer students from out of state or simply missing data from the previous year. Simple analysis of students with only a single year of test data revealed no obvious systematic missing data issues.

cation level into those who have a high school education or less and those who have above a high school education. Students in charter schools and alternative schools are dropped.

The other unique feature of the data set is that it contains teacher absence data. Teacher absences in North Carolina are broadly categorized into sick leave, personal leave, and annual (vacation) leave. I use the sum of sick leave and personal leave as the measure of teacher absence in an academic year. Most of the annual leave days coincide with school vacation days. About 60 percent of annual leaves are concentrated in December, June, July, and August. Another 20 percent occur in November, which indicates Thanksgiving break. Most of the remainder of the annual and personal leaves are for months at a time, indicating longer absences due to severe health problems or maternity leave.

Because sick and personal leaves are unplanned for and take place during the school year, they can have the largest impact on student learning. A substitute teacher may have less experience and lack certification. More importantly, he or she may have a weaker grasp of the material being taught, will not know the aptitude of individual students, and may not be induced to exert more effort from bonus incentives. For the study period, teachers were absent an average of 9.6 days in an academic year of 180 days. There is a significant number of teachers who are absent more than 30 days (about 5 percent of the original sample), and I exclude these teachers from the analysis. The resulting teacher absence rate of 5.3 percent is in line with other studies of teacher absences (see Ehrenberg et al. 1991). Including teachers with a high level of absence does not qualitatively change results. Table 1 summarizes student, teacher, class, and school characteristics.

V. Econometric Model

The econometric model estimates three equations, following the theoretical model. School-level expected bonus is generated using student achievement and the accountability rules defined in Section II. Teacher effort is captured using teacher absence. I show that teacher absence can be separated into effort and other unrelated shocks and use the incentive legislation to isolate the effort component. This reduces noise arising from uncorrelated shocks, making teacher absence a good signal for effort. Student achievement is estimated using distilled effort with other traditional student, peer, and teacher demographic characteristics in a standard education production function. The system is solved iteratively until convergence.[22]

[22] This is equivalent to solving for a teacher reaction function. An alternative specification of estimating student achievement nonparametrically using student, peer, teacher, and school data yielded no qualitative differences.

TABLE 1
SAMPLE STATISTICS

Variable	Mean	Standard Deviation	[Min, Max]
Student ($N = 897{,}471$):			
Male	.504		
Minority	.346		
Parent high school or less	.724		
Gifted	.137		
Disabled	.114		
Limited English	.022		
Teacher ($N = 45{,}011$):			
Male	.077		
Minority	.153		
Experienced	.942		
Certified	.056		
Absent days/academic year	9.562	6.107	[0, 30]
Class:			
% male	.511	.093	[0, 1]
% minority	.353	.280	[0, 1]
% parent high school or less	.703	.263	[0, 1]
Class size	22.01	4.151	[5, 36]
School:			
School size	517.43	196.199	[46, 1,423]
Number of teachers	34.43	11.944	[4, 77]
Rural school	.468		

Source.—North Caroline Education Research Data Center data set, years 1999/2000–2003/4.
Note.—Included are all third- to fifth-grade public elementary students and their teachers, excluding charter and alternative schools. Teachers with more than 30 days of absence are excluded. Students with zero or one exam record are excluded. Experienced equals one for teachers with more than 1 year of experience because of limitations of the administrative data set in tracking total years of employment.

A. *School-Level Bonus*

While the expected bonus would usually be estimated as a probit or logit regression from data on current-year bonus receipt, last-year test scores, and this-year predicted scores times the bonus amount, I am able to use the state-defined criterion as specified in the accountability system description. Because the differences between school scores and state criterion are standardized to $N(0,1)$, let $\hat{Y}_{st}$ represent the predicted school performance this year, and let $Y_{s(t-1)}$ be school performance last year.[23]

Expected bonus is

$$E(B) = [\Phi(\hat{Y}_{st} - Y_{s(t-1)} - Cr_1) + \Phi(\hat{Y}_{st} - Y_{s(t-1)} - Cr_2)] \cdot B, \qquad (4)$$

[23] The Y values are the sum of all students' standardized reading and math scores. The term $Y_{s(t-1)}$ captures last-year test scores of all current students in school s, even if they were not attending school s last year.

where Cr_1 and Cr_2 are expected and high-growth composite threshold values, respectively, B is the bonus amount, and $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF).[24]

B. Teacher Absence Decision

If effort were observable, the effect of incentives could be captured by running the following regression:

$$e_{jst} = \alpha_0 + X_{jst}\alpha_1 + \alpha_2 I_{jst} + \varepsilon_{jst},$$

where e_{jst} is teacher j's effort (located in school s in year t); X_{jst} are observable characteristics of teacher and year, school district, and grade dummy variables; I_{jst} are measures of incentive strength (bonus receipt probability and incentive dilution); and ε_{jst} is the idiosyncratic error.

While effort is not directly measurable, teacher absence can serve as a noisy signal. Teacher absence is noisy because it contains effort as well as unrelated shocks and measurement error. I hypothesize that teacher absence, A_{jst}, is determined by three components,

$$A_{jst} = g(X_{jst}, e_{jst}, \eta_{jst}),$$

where η_{jst} represents factors that affect absence that are unrelated to e_{jst}, such as unforeseen bad/good health outcomes and weather, and other orthogonal shocks that affect the number of days a teacher is absent.

I assume that for two teachers i and j, $E(A_i) \geq E(A_j)$ if and only if $e_i \leq e_j$ conditional on teacher, class, school, and district characteristics.[25] Because the absence variable is a count of days of absence in the school year, it is appropriate to think of effort as the aggregate effort provided by the teacher throughout the academic year. This assumes away situations in which a teacher may redistribute effort for different time periods. For instance, a teacher cannot "save up" her effort prior to the exam and saturate her students at exam time.[26]

Since absence is correlated with effort and incentives are correlated with effort, the projection of incentives on absence results in an unscaled measure of effort. Using the data available and the first-order condition directly from the theory, I model the absence decision as follows:

$$A_{jst} = X_{jst}\alpha_1 + \alpha_2 \ln J_{st} + \alpha_3 \ln[\phi(\hat{Y}_{st}, Cr_1) + \phi(\hat{Y}_{st}, Cr_2)] + \eta_{jst}. \quad (5)$$

[24] In the North Carolina case, $B = 750$ as the expected growth yields \$750 and high growth yields \$1,500. Note from the description of the accountability system that the argument inside $\Phi(\cdot)$ is a z-score distributed with a zero mean and a standard deviation of one. More generally, the equation would be

$$E(B) = [\Phi(\hat{Y}_{st} - Y_{s(t-1)} - Cr_1)B_1 + \Phi(\hat{Y}_{st} - Y_{s(t-1)} - Cr_2)]B_2.$$

[25] For instance, it is possible that experienced teachers may need to exert less effort compared to a new teacher for the same level of education production.

[26] See App. A for a formalization of this condition. Also, teacher absence pattern does not fluctuate just prior to the end of the school year.

The measures of incentive strength (I_{jst}) to be used are J_{st}, the number of teachers at the school, and $2\phi(\hat{Y}_{st}, Cr_1) + \phi(\hat{Y}_{st}, Cr_2)$, the position of the school with respect to the probability of bonus receipt.[27] The J term estimates free-ridership, as discussed in the theory. I anticipate $\alpha_2 > 0$. The effect of the ϕ term is best explained by figure 3. The ϕ term measures the incentive effects of being at different points in the distribution of bonus receipt probability. The shape of the log of the pdf is strictly concave, with low effort tied to very low and very high probabilities of bonus receipt. The peak of effort exertion is tied to midlevel academic growth, when the bonus receipt is in doubt, and is very dependent on the effort exertion of teachers. Since absence is negatively associated with effort, I expect $\alpha_3 < 0$. I assume that the teacher does not differentiate effort exertion between subjects or among students.[28]

The shape of the ϕ variable is important as it reflects the nonlinear effect of the probability of receiving the bonus. As the bonus becomes very easy or very difficult to attain, the utility-maximizing response of teachers is to decrease effort.[29] The term X_{jst} is teacher, class, and school characteristics;[30] η_{jst} is the idiosyncratic error term that represents the uncorrelated health and weather shocks.

By estimating the absence decision in this way, I decrease the noise in teacher absence and distill it into a useful signal for effort. The change in "predicted absence," driven purely by the incentives in the bonus program, is an indicator of change in effort because absence reduction per se is not rewarded as part of the incentive system.[31] Predicted absence, $\hat{A}$, is now a latent variable for effort.

Distilling effort in this way differentiates and isolates its impact on student achievement from other possible ways of increasing student achievement. Many of the methods mentioned in the introduction (such

[27] I use the total number of teachers at the school (which includes teachers in kindergarten, first, and second grades) to account for the fact that teachers who do not contribute to the school average score are in line to receive the cash bonus. Using the count of only third-, fourth-, and fifth-grade teachers does not change the qualitative results.

[28] Implicit in the equation, by the exclusion of teacher fixed effects, is the strong assumption that there is no unobservable heterogeneity of teachers that would result in different patterns in teacher absence. This was primarily due to data problems in being able to consistently link teachers across the entire sample.

[29] The nonlinear, nonmonotonic effect will be important for policy analysis. To check if the shape of the pdf is consistent with the theory, I run an alternative specification in which I replace ϕ with $E(B)$ and $E(B)^2$. The squared term of expected bonus is included to capture nonlinear effects of the probability of receiving the bonus. Results support the functional form assumption.

[30] These include teacher gender, minority, experience, certification status, class and school size, percentage minority students, percentage low parent education, a rural school indicator, and dummy variables for grade, district, and year.

[31] That absence is not part of the incentive system is important for its role of serving as a signal for effort. If the number of absences was incorporated into the incentive system, depending on the state formula, teachers could choose to simultaneously reduce effort and absence, while ensuring the same or higher expected payout. This would render absence less useful as a signal. For instance, in the system analyzed by Imberman and Lovenheim (2012), teacher absence does affect the bonus.

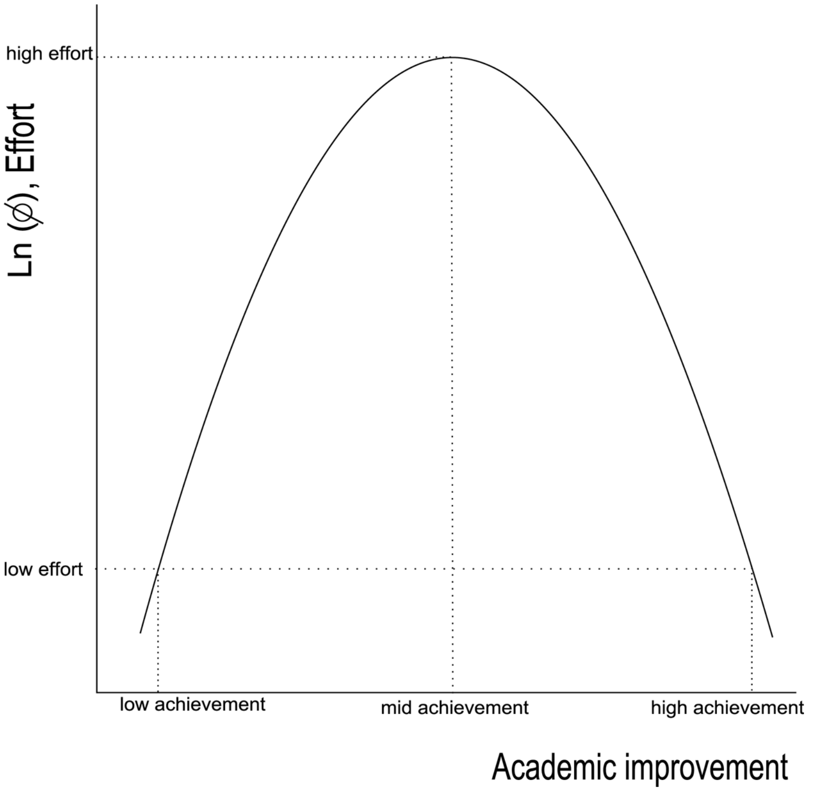

Figure 3.—The effect of ϕ on effort

as excluding marginal students from taking the exam or changing answer sheets) serve as substitutes for education production through effort.[32] If test score gains are observed for schools under a high degree of accountability pressure yet distilled effort response is absent, this may indicate that schools are employing other methods.

C. Student Achievement

The student achievement equation does not follow from any specific utility maximization solution for students. I assume that there exists some student-level production function for education in which the inputs are student, peer, and teacher characteristics. The achievement function for student i in subject m is specified as follows:

$$y_{mist} = Z_{ist}\beta_1 + \beta_2 y_{mis(t-1)} + \beta_3 \hat{A}_{jst} + \beta_4 \hat{A}_{jst} \cdot I[\text{minority}] + \nu_{mist}, \qquad (6)$$

[32] If teachers decided to exert more effort and increase the use of these other techniques at the same time, the impact of effort would be overestimated in the model.

where Z_{ist} is a vector of student, peer, and teacher demographic characteristics plus year, district, and grade dummy variables;[33] $\hat{A}_{jst}$ is the projected measure of (negative) effort from equation (5); I[minority] is an indicator variable that equals one if the student is black or Hispanic; and the interaction between effort and minority status was included to check if "per-unit" effort efficacy differed by demographic characteristics.

D. Estimation

Equations (4), (5), and (6) are solved iteratively until convergence. In the initial step, I plug in the actual test scores from the data, Y_{st}, in equation (4) to generate the probability of qualifying for the bonus, $\Phi(\cdot)$, and expected bonus.[34] Equation (4) is a rule set, not a regression.[35]

The CDF is converted to the pdf $\phi(\cdot)$ for use in the absence equation (eq. [5]). The $\phi(\cdot)$ value, observable teacher and school characteristics, as well as the measure of free-ridership (J) are used to generate projected effort ($\hat{A}$). Projected effort is then plugged in the student achievement equation (eq. [6]) with student, teacher, and school characteristics to predict student test scores. Predicted student test scores are aggregated to school averages, $\hat{Y}_{st}$, and plugged into equation (4) again to generate new expected bonus and $\Phi(\cdot)$, and the process continues until the parameters from equations (5) and (6) converge, essentially achieving a fixed point in the parameter and the distribution of test scores.[36]

VI. Results

Having specified the estimation strategy, I present the results. Before the parameter estimates are considered, it is natural to wonder if the procedure to distill effort is required at all. A single-equation ordinary least squares (OLS) or fixed-effects model that uses raw absence data may suitably summarize the relationship between teacher effort and achievement. Alternatively, it is possible that effort is not important. Pure exposure to teachers may drive achievement.

[33] Student characteristics are gender, minority, parental education status, and the previous year's test score. Peer characteristics are percentage minority, percentage low parental education (at class and school levels), and class and school size. Included teacher characteristics are teacher gender, minority, experience, and certification status.

[34] Expected bonus is used for simulation and cost-benefit analysis of potential policy changes. An assumption inherent at the start of the estimation is that the actual test scores serve as a good initial proxy for the average test score before the shock in eq. (2) is realized. This is more likely to be an acceptable assumption if schoolwide shocks tend to be small in magnitude in either direction.

[35] The actual bonus would equal \$0 if $Cr_1 > Y_{st} - Y_{s(t-1)}$, \$750 if $Cr_1 \leq Y_{st} - Y_{s(t-1)} \leq Cr_2$, and \$1,500 if $Y_{st} - Y_{s(t-1)} > Cr_2$. The math and reading test scores of students in grades 3, 4, and 5 are condensed into a z-score that measures schoolwide achievement, which directly affects the teachers' incentives, which in turn affects student scores.

[36] An alternative interpretation of the setup is a simultaneous equations system with the probability of bonus receipt, number of teachers, and student individual characteristics serving as the exclusion restrictions, with effort/absence and education achievement as the endogenous variables.

TABLE 2
OLS Estimates of the Impact of Absence on Academic Achievement

Variable	Reading Coefficient	Reading Standard Deviation	Math Coefficient	Math Standard Deviation
Absence	−.0010	.0001	−.0022	.0001
Absence × minority	.0001	.0002	.0002	.0002
Last-year subject score	.7085	.0007	.7454	.0007
Male	−.0378	.0013	−.0060	.0012
Minority	−.1693	.0027	−.1437	.0025
Parent high school or less	−.1887	.0017	−.1737	.0016
Teacher male	−.0259	.0025	−.0204	.0023
Teacher minority	−.0053	.0020	−.0338	.0019
Teacher certified	.0290	.0028	.0500	.0020
Teacher experienced	.0774	.0029	.1134	.0027
% class with parent high school or less	−.0078	.0049	−.0331	.0046
% class minority	−.1125	.0074	−.1306	.0070
% school with parent high school or less	−.1190	.0066	−.0405	.0062
% school minority	.0555	.0083	.0690	.0073
Class size	.0036	.0004	.0035	.0004
School size/1,000	−.0267	.0169	.0119	.0158

Note.—Estimation included district, year, and grade fixed effects. Dependent variables are standardized EOG reading and mathematics exam scores.

To motivate the need for signal distillation, I start by showing OLS estimates without isolating the effort component in the first step by estimating equation (6) using raw absence data. The results are shown in table 2. The negative impact of teacher absence on academic achievement is similar to results from Clotfelter et al. (2007a).[37]

If the regression results are interpreted as causative, students do not suffer much from teacher absence. Enacting some policy that would cut the average number of absent days in half, from approximately 10 days to 5 days, would increase average reading achievement by roughly 0.5 percent of a standard deviation and math achievement by about 1 percent of a standard deviation. In contrast, paying for teacher certification would be approximately three times more effective. The OLS estimate may be small because an absence does not result in zero education production during that day. A substitute teacher will be assigned, and students will still receive instruction. Imagine two identical teachers teaching identical classes. If one of the teachers is absent an additional day compared to the other teacher (with a substitute teacher filling in), student performances should not be significantly different across the two classes.

The estimate of the impact of projected effort due to incentive effects in table 3 is more than 25 times greater than the OLS parameter estimate on absence for reading performance and more than 15 times larger for math performance. However, as I show later, this magnified impact still

[37] Clotfelter et al. (2007a) show that fixed-effects estimates yield smaller but still significant estimates on absence. The reading and math scores are adjusted to represent a percentage change of one standard deviation of scores.

TABLE 3
STRUCTURAL ESTIMATES OF THE IMPACT OF EFFORT ON ACADEMIC ACHIEVEMENT

	Reading		Math	
Variable	Coefficient	Standard Deviation	Coefficient	Standard Deviation
Absence	−.0250	.0052	−.0342	.0048
Absence × minority	.0023	.0007	.0018	.0007
Last-year subject score	.7086	.0007	.7456	.0007
Male	−.0378	.0013	−.0063	.0012
Minority	−.1900	.0074	−.1599	.0069
Parent high school or less	−.1884	.0018	−.1734	.0016
Teacher male	−.0555	.0071	−.0604	.0066
Teacher minority	.0044	.0030	−.0208	.0028
Teacher certified	.0667	.0088	.1006	.0082
Teacher experienced	.1803	.0231	.2535	.0216
% class with parent high school or less	−.0065	.0049	−.0308	.0046
% class minority	−.1125	.0075	−.1307	.0070
% school with parent high school or less	−.1091	.0069	−.0267	.0065
% school minority	.0705	.0090	.0901	.0084
Class size	.0033	.0004	.0032	.0004
School size/1,000	−.0846	.0216	−.0655	.0020

Note.—Estimation included district, year, and grade fixed effects. Dependent variables are standardized EOG reading and mathematics exam scores.

translates to modest gains in test scores. This difference can be explained by extending the example above. Assume that one teacher is more motivated. The less motivated teacher does not prepare as much and generally cares less about the educational outcome of students. Further assume that the two teachers are in charge of identical classes. All else equal, student achievement in the less motivated teacher's class should be lower, and this teacher is likely to be absent more often throughout the school year. Now, assume that the enthusiastic teacher has a bad health shock and is forced to be absent the same number of days as the less motivated teacher. While students in both classes are exposed to their teacher the same number of days, the enthusiastic teacher will give superior instruction to her students throughout the entire academic year. The OLS procedure in this case treats the teachers as identical (because of the same number of absences) and predicts the same achievement (the average effect across the two teachers). The structural model predicts higher effort and achievement in the motivated teacher's class due to the incentives from the accountability system.

In addition, teacher certification status and teacher experience become stronger predictors of academic achievement when absence is distilled into effort. Teachers with more experience also have higher rates of absence (perhaps as a result of more health issues or smaller valuation of the bonus; see table 4). The full impact of experience on academic achievement is shared between the teacher experience and the raw absence variables. This pushes the OLS parameter estimate on experience down and pulls the parameter on absence up. Since the parameter estimate on absence is a negative number, the true impact of teacher absences is fur-

TABLE 4
Teacher Effort Decision Estimates

Variable	Coefficient	Standard Deviation
Log number of teachers	.3858	.1561
Log ϕ	−.2613	.1067
Male	−1.2871	.1561
Minority	.4215	.0843
Experienced	4.4572	.1254
Certified	1.6172	.1188
Class size	−.0825	.0351
School size/1,000	−3.7188	.8637
% class with parent high school or less	.6376	.4948
% class minority	.1411	.2798
% school with parent high school or less	−.0068	.5327
% school minority	.4109	.3189

Note.—Dependent variable is days of teacher absence. Estimation included district, year, and grade fixed effects.

ther underestimated in this framework.[38] Using the absence response of experienced teachers to changes in expected bonus separates out this confounding influence.[39]

The impact of effort is some 30–40 percent greater for the math test. In addition, teacher characteristics such as experience and certification status have a much larger impact on student scores for math compared to reading. This outcome is in line with the literature, which usually finds that teachers are more effective at raising math scores (see, e.g., Chetty, Friedman, and Rockoff 2011). The impact on effort (predicted absence) is also similar to that estimated by Clotfelter et al. (2007a) in that absence has a greater impact on math scores. A reason for this difference in effectiveness may be that parents have more influence on students' reading skills (especially when they are young), whereas most students learn math in school.

The teacher effort estimates in table 4 present the impact of incentive forces on teacher absence. The probability of qualifying for the bonus, represented by ϕ, and the number of teachers in the school, J, change the level of effort teachers exert in response to the accountability system.

The sign of the parameter on the ϕ term shows that the school's location on the pdf of the bonus receipt probability is critical in determining a

[38] When the OLS regression is run excluding the teacher experience and certification variables, the parameter estimate on absence increases from −0.0010 to −0.00045 while other parameter estimates are stable. This further confirms that the raw absence variable is soaking up some of the explanatory power of experience.

[39] Of note is that the parameter on class size is positive in the student achievement equation. This seems to imply that student achievement increases by roughly 0.3 percent of a standard deviation for each additional student introduced to the class. This may be indicative of sorting, with principals assigning more effective teachers to larger classes. Another possibility is that the increase in class size is distorting the incentive effects. A teacher with a larger class necessarily has a larger impact on the bonus outcome. Some of the increase in effort may be picked up by the parameter on larger class size. See Sec. III for a more detailed discussion.

teacher's effort exertion. Very low and very high probabilities of bonus receipt are associated with lower effort exertion. With the ϕ term translated into expected bonus, at low levels of expected bonus, teachers exert low effort, with effort increasing as expected bonus (probability of bonus receipt) increases. The peak level of effort exertion is achieved when expected bonus is approximately \$870. Beyond this point, the bonus becomes a "sure thing," and effort declines again toward lower levels as expected bonus increases. Therefore, if the goal of the incentive policy is to energize teachers to exert the maximum amount of effort possible, the threshold value for attaining the bonus must be set such that qualifying for the cash bonus is neither too easy nor too difficult. While the estimate for effort signal shows how teachers respond to incentives, the estimate for the effect of effort on achievement shows how students respond to motivated teachers. Increasing the expected bonus amount from \$0 to \$850 will increase an average teacher's effort by about 13 percent, which translates to about a 2.8 percent of a standard deviation increase on the average student's EOG reading exam and about a 3.8 percent of a standard deviation increase on the mathematics exam. It is interesting to place these results in the context of the effectiveness of class size reduction (see Rivkin et al. 2005). While estimated magnitudes differ by studies, it is generally agreed that cutting a 20-student class in half can yield somewhere between a 5 percent and a 15 percent of a standard deviation increase in student achievement. In a comparison of the effectiveness of the two policies, bonus incentives may be an efficient method of raising students' achievement.

The parameter on the J term shows that a teacher exerts a lower level of effort when the number of colleagues in the school increases. A higher number of teachers at a school implies lower effort for teachers, pointing to incentive dilution in school-level incentives.[40]

A. *Identification*

Identification in this model is not based on pre- and posttreatment outcomes. The data set does not go back in time far enough to capture pretreatment outcomes. Only about 25 percent of the student data can be matched across any two preincentive years.[41] Instead, identification comes from the variation in the number of teachers and the distance from the cut-

[40] A possible concern here is that the parameter on the number of teachers is not measuring free-ridership arising from the bonus regime but a preexisting condition. I ran a regression that measures the sensitivity of absence on the number of teachers in the school for pre- and postincentive years. The parameter estimate on the number of teachers preincentive is statistically insignificant, while it is positive and significant for parameters from postincentive installation.

[41] Given this limitation, I run a placebo trial that estimates the teacher absence decision equation and the effect of incentives on achievement. The results show that teachers do not respond and student achievement is not driven by incentives (as they do not exist yet). There is some evidence that once the bonus was discontinued, the "bonus effect" disappeared. The bonus system was discontinued in 2010 as a result of budgetary cuts. See Ahn and Vigdor (2012) for details.

off for a bonus in both directions. There are some schools that are so far below the standard across the sample years that their predicted probability of qualifying for the bonus is near zero, and there are some schools that are far enough over the bar on all years that the bonus is almost a certainty. The relatively high number of days of absence observed for teachers in these schools is what drives identification in the model. A necessary assumption here is that teacher effort is fully captured by observed teacher absence. If teachers exert effort on unobservable dimensions and if this effort exertion is uncorrelated with absence (and/or incentive pressure), the parameter estimate on projected effort will be biased. If, on the other hand, these dimensions of effort move together with absence, projected effort can be interpreted as the combination of many types of teacher effort.

Incentive-related factors that move teacher effort in the model are captured by the number of teachers at the school (J_{st}) and the position of the school with respect to the probability of bonus receipt ($\phi(\hat{y}_{st}, Cr_1) + \phi(\hat{y}_{st}, Cr_2)$). In essence, J and ϕ instrument for teacher effort. The econometric model is identified if the instruments J and ϕ can be excluded from the student achievement equation (eq. [6]); however, there are credible arguments for including these exclusion restrictions directly into the student achievement equation.

The number of teachers, J, serves as a measure of incentive dilution from free-rider effects that could affect the amount of effort teachers put forth. If the free-rider effect leads to more absences, this could negatively affect student achievement. However, it is also possible to argue that more teachers at a school could lead to more individualized attention or better discipline of students. Further, J as a measure of free-ridership and J as a measure of individualized attention could pull achievement in opposite directions. To test whether J has a direct positive effect on student achievement, I run an alternative specification using $J \cdot I[\text{no bonus effect}]$ and find that there is no statistically significant direct effect of J on y. Here, $I[\text{no bonus effect}]$ is an indicator variable that equals one when the probability of obtaining the bonus is near 100 percent or 0 percent and zero otherwise. That is, the bonus is almost assured or virtually out of reach, meaning that the effect of J should come only through the direct channel. See table 5.

A school's proximity to the bonus threshold may also affect student achievement through other noneffort channels. For example, if the school is close to the threshold, the principal or superintendent may decide to reallocate the school budget toward expenditures that could bring about an immediate short-term increase in student achievement (perhaps at the cost of longer-term expenditure goals). Alternatively, money could be reallocated in schools that are farther away from the threshold in an attempt to bring them closer to the middle of the distribution, where the incentive effect will be strongly felt by teachers. It may also be possible to argue that the bonus itself directly affects teacher behavior unrelated to

TABLE 5
Testing for Direct Effect of Teacher Exposure on Achievement

Variable	Coefficient	Standard Deviation
Log absence	−.2452	.0090
Log absence × minority	.0148	.0057
Log number of teachers × Pr(bonus) ≅ 0 or 1	−.0029	.0035
Last-year reading score	.7104	.0007
Male	−.0372	.0013
Minority	−.1866	.0130
Parent high school or less	−.1882	.0018
Teacher certified	.0700	.0031
Teacher experienced	.2434	.0066
% class with parent high school or less	−.0415	.0040
% class minority	−.0901	.0040
Class size	−.0010	.0002
School size/1,000	−.0036	.0006

Note.—Estimation included district, year, and grade fixed effects. Dependent variable is standardized EOG reading exam scores.

the time commitment component of effort as revealed by absence. For example, a teacher may upgrade (or choose to skimp on) teaching aids on the basis of anticipated payouts. Table 6 tests for these alternative channels by including dummy variables for schools that are almost out of the bonus and completely assured of the bonus. The insignificant parameter estimate on I[no bonus effect] show that the bonus amount has no direct effect when predicted effort is included in the regression.[42]

The two regressions show that the two measures serve as valid exclusion restrictions. The number of teachers in a school is a good measure of free-rider effects, and the location of the school along the probability distribution serves as a good proxy for incentive effects.

B. Impact of Sorting

Student achievement gains in response to accountability pressure are assumed to be captured by the change in absence response of teachers in the model. However, there may be other school responses that occur concurrently with teacher effort increase. If these responses also help to increase student achievement, the effectiveness of effort may be overestimated. A central assumption is that possible nonrandom sorting and matching of teachers across and within schools in response to accountability pressure will not significantly bias the estimation in the student achievement function. If schools and teachers respond to accountability

[42] While the North Carolina data do not contain school-level budget information, they do contain district-level (local education authority) expenditure data. A regression with various nonsalary current expenditures (at the district level) with the fraction of schools near the threshold as the relevant control failed to yield statistically significant estimates. While this regression may fail to isolate in-year changes or details that would be visible with school-level data, it does support the assumption that no large-scale budgetary reallocation is taking place in response to the bonus incentives.

TABLE 6
Testing for Direct Effect of Bonus on Achievement

Variable	Coefficient	Standard Deviation
Log absence	−.0165	.0067
Log absence × minority	.0028	.0007
I(<5% prob. of bonus receipt)	.0028	.0048
I(>95% prob. of bonus receipt)	−.0067	.0044
Last-year reading score	.7035	.0007
Male	−.0384	.0013
Minority	−.1998	.0071
Parent high school or less	−.1905	.0017
Teacher certified	.0487	.0112
Teacher experienced	.1352	.0298
% class with parent high school or less	.0256	.0046
% class minority	−.1755	.0075
Class size	−.0006	.0002
School size/1,000	−.0008	.0002

Note.—Estimation included district, year, and grade fixed effects. Dependent variable is standardized EOG reading exam scores.

pressure by sorting, the gains to effort may be distorted. Teacher sorting may occur in the short, medium, and long term.

Over the short term (within the current academic year), a principal may seek to match students and classes to particular teachers to maximize the probability of bonus receipt. There is little doubt that nonrandom sorting of students and teachers into classrooms exists.[43] This also holds true in the North Carolina data; however, this is problematic only if sorting occurs more frequently in schools that are under accountability pressure, compared to schools not under pressure.

If there is substantial sorting of teachers on unobservable characteristics in response to accountability pressure within the school, the impact of effort may be overestimated. Since the observed increase in test scores in pressured schools resulted from the additional effort induced by the incentives plus the sorting and since sorting and effort increase are correlated (via accountability pressure), the impact of effort will be biased upward. This bias due to sorting on unobservable characteristics may be of some concern if the unobservable (positive) teacher characteristics are not homogeneously distributed across all schools along the bonus CDF. Schools with more teachers with desirable unobservable characteristics may have the ability to more effectively match teachers to students to maximize education production. For instance, if the proportion of teachers with desirable unobservable characteristics is positively correlated with the school's location on the bonus CDF, a policy change that moves schools on the left and right tails to the middle of the distribution may yield larger than expected gains for schools that were at the right tail and

[43] There is some evidence that students and teachers paired by race (e.g.) tend to do better (see Dee 2004). See Clotfelter, Ladd, and Vigdor (2006), among others, for examples of nonrandom sorting.

smaller than expected gains for schools that were at the left tail, provided that both types of schools sort to take advantage of the unobservable teacher characteristics.[44] The literature has been agnostic about why sorting occurs, but many studies show that teachers with more seniority are often in classrooms that have a larger proportion of higher-achieving students. If most sorting occurs because of seniority rules or personal and professional relationships with the principal and not because of accountability pressure, the bias is not expected to be large.[45]

Over the medium term (across 1 or 2 years), teachers may respond to accountability pressure by attempting to transfer to schools that offer a higher likelihood of qualifying for the bonus. Alternatively, schools under higher pressure may seek out teachers who can increase achievement.[46] The decision to transfer next year or in the near future should not affect a teacher's effort in the current academic year since the probability of bonus receipt is determined year to year. It is possible to argue that a teacher may seek to maximize effort output this year as an "audition" to transfer to a preferred school next year. In this case, the incentive effect may be underestimated since teachers in schools that have a low probability of qualifying for the bonus may exert additional effort in order to facilitate a transfer to a better school in the near future.

Over the long term (across several years), if the bonus is perceived as a pay increase, the state may be able to attract a larger number of higher-ability teachers. As Hanushek (2009) demonstrates, replacing even a small fraction of low-ability teachers with average teachers may have a large positive long-term impact. If retiring teachers are continually re-

[44] It should be noted that schools located toward the left tail of the CDF are not necessarily "undesirable schools" in the traditional sense (low test scores, urban, poor, etc.). In fact, the correlation between a school's ABC status (whether it qualifies for the bonus) and whether the school made AYP is relatively small. Even if there is high correlation between positive unobservable teacher characteristics and the desirability of the school (as measured by AYP), the correlation between the positive unobservable teacher characteristics and her school's place on the probability CDF may not be as high.

[45] To test for possible sorting of observable characteristics in response to accountability pressure, I ran the following regression:

$$\text{DevMin} = \beta_0 + \beta_1 \cdot \text{pdf} + \beta_2 \cdot \text{tchexp} + \beta_3 \cdot (\text{tchexp} \times \text{pdf}) + z + \epsilon,$$

where DevMin is classroom minority percentage divided by school minority percentage (to capture overt redistribution of students), pdf is the measure of accountability strength, tchexp is the teacher experience variable, and (tchexp $\times$ pdf) is the interaction term. The vector z contains all other variables, including teacher demographic characteristics and indicators for year, grade, and district; ϵ is the error term. The β_2 term is negative and significant, confirming that sorting on seniority does occur. The regression results are insignificant for β_3, lending weight to the idea that the sorting is not occurring as a result of accountability pressure. The intuition is that if the classroom composition is a policy tool (and principals want to use this tool by matching effective teachers with either high or low minority-heavy classes to maximize education production), as pressure increases, we should see a positive β_3 if effective teacher high-minority class is attempted and negative β_3 if effective teacher low-minority class is attempted. Note that this type of test can check for sorting only on observable teacher characteristics.

[46] In Ahn (2013), teachers with higher fixed-effect values tend to match more frequently with schools that are under more accountability pressure.

placed with higher-quality teaching candidates on average, statewide academic achievement may increase. The ABC bonus program began paying out in 1996/97. The starting year of the data set used in this study is 1999/2000. It is possible that some on the fence about entering the teaching profession could have been enticed to do so because of the bonus. If better teachers enter the labor market, the "true" impact of the bonus may be underestimated. The true impact would be the incentive effect plus the increase in the "baseline" academic achievement of students due to the increase in the average quality of teachers.

Students sorting across schools in response to accountability pressure may also distort the accountability system. Students moving to (or away from) schools under pressure will not necessarily raise or lower a school's bonus receipt probability. Because previous-year test scores are attached to the student and not to the school, the test score growth requirement will remain stable even with student transfers. However, students may affect classmates' test scores through peer effects, and movements of a large number of students in and out of schools may be disruptive not only for the transferring students but for their old and new classmates as well. In the data, the transfer rate of students is not strongly correlated with accountability pressure. If students with higher test score growth potential are more likely to transfer, this may lead to incentive effects being overestimated.[47]

While the ability to measure teacher effort and its effect on student achievement are interesting in their own right, the more important question is how to design policy to effectively induce effort. The next section presents two possible policy changes. The first experiment changes the criterion from schoolwide performance to a classroom-level performance measure, attempting to eliminate the incentive dilution problem. The second experiment changes the current standard to a grade-level performance measure.

VII. Policy Simulations

An incentive system that rewards on the basis of group performance may suffer from free-rider problems. Therefore, targeting the bonus at the individual level may improve performance. I present two possible policy alterations to the current school-level rewards to examine possible effects. First, I change the accountability criterion to target bonus at the classroom level, completely eliminating the free-rider effect. Second, I change the criterion to target at the grade level, an intermediate level of grouping between school level and classroom level.[48]

[47] The student takeup rate for transfers when schools failed to make adequate yearly progress under NCLB in North Carolina was also very low during the sample years.

[48] It is worth noting that a continuously increasing individual-level bonus payment in student-level achievement growth (assuming revenue neutrality) would almost surely be the most efficient policy change. The simulations conducted in this section are more modest al-

Computationally, I change the bonus rule (eq. [4]) to reward teachers at either the grade level or the classroom level by changing the level at which student scores are aggregated. This generates a new probability of bonus receipt for each teacher. The new probability and the appropriate count of teachers (one for classroom level or the appropriate number of teachers for grade level) are plugged into equation (5), keeping the parameter values fixed. This generates a new effort level (absence), which is then plugged into the student achievement equation (eq. [6]), again, with the estimated parameter values fixed. The newly generated test score outcomes are summed and averaged according to the new bonus rule and plugged into equation (4), generating a new probability of bonus receipt. The system is iterated until the average difference in each teacher's absence from one iteration to the next is less than 1 percent.

Classroom-level incentives simulation are presented in column 2 of table 7.[49] The results show that the expected bonus per teacher increases by about $150. Since there are roughly 45,000 teacher/year observations over the 5-year sample period and assuming that there is an equal number of kindergarten, first-, and second-grade teachers in the qualifying schools, gross bonus payout would increase by roughly $13.5 million each year. This is approximately equivalent to a 13 percent increase in the state budget for bonus payout. Yet at the same time, average predicted effort decreases by about 5.6 percent and the variance of expected bonus (and by extension variance of effort) increases.

These seemingly contradictory results can be explained by examining teacher incentives. With the regime switch from school level to classroom level, teachers would be split sharply into those who can achieve the bonus standard and those who cannot. Going from a schoolwide effort to an individual effort will starkly separate low- and high-ability teachers.[50] A teacher with a low chance of success will have little incentive to exert effort, while a teacher with a fair chance of success will increase her effort only until the marginal benefit of doing so matches marginal effort cost. In addition, teachers with very high ability and/or very high-achieving students will decrease effort since the probability of bonus receipt is no longer dragged downward by lower-achieving classes. With the right and left tails of the expected bonus distribution pushing effort downward, average effort decreases.

terations in line with the types of policy changes that a legislative body may be willing to consider. For example, in 2005/6, the rule set for high school bonus calculations was amended to take into account two previous years' test scores while keeping the same payout structure.

[49] It is possible that as incentive structures change, classroom allocations of students may also change. For instance, a principal may optimally reallocate students to get all teachers to the middle of the bonus probability to induce greater effort (see Barlevy and Neal 2012). The simulations do not account for this possibility.

[50] "Low" ability does not necessarily indicate poor-quality teachers, but teachers with a low probability of success. This could be the result of the composition of students in her class or school, e.g.

TABLE 7
CLASSROOM AND GRADE-LEVEL TARGETING

	School (1)	Classroom (2)	Grade (3)
Expected bonus	737.69	889.56	858.64
	(149.69)	(562.96)	(456.11)
Effort increase (%)		−5.62	+3.70
Δ reading score (%)*		−1.04	+1.32
Δ math score (%)*		−1.38	+1.72
Teachers with higher effort (%)		35.0	92.7
Schools with higher scores (%)		46.5	88.0
Students with higher scores (%)		33.0	93.2

* Change in percentage of one standard deviation in test scores.

Therefore, moving from schoolwide to classroom standards eliminates the direct negative effect of free-ridership but is offset by locating more teachers at the tails of the bonus pdf. The histograms in figures 4, 5, and 6 demonstrate this indirect effect. Figure 4 shows that switching the bonus criterion to classroom performance increases the weight of the tails of the expected bonus distribution. The right tail represents classrooms that were outperforming the rest of the school yet were being dragged down as part of the schoolwide criterion. The left tail represents classrooms that were underperforming compared to the rest of the school yet were buoyed by being part of the schoolwide criterion. Figure 5 shows that as the bonus criterion changes, teachers adjust their efforts accordingly.[51] While there is little change in the high-effort (low-absence) part of the distribution, the percentage of teachers exerting low effort (high-absence) increases. This is directly attributable to the increase in the tails of the histogram in figure 4. Highly under/overachieving teachers are no longer pushed toward the middle of the distribution, resulting in lower effort exertion for these teachers. The decrease in average teacher effort is reflected in the increase in the fraction of students with lower standardized scores and the slight decrease in students with higher standardized scores as illustrated in figure 6.

If attempting to eliminate the free-rider effect through individual incentives results in negative achievement growth due to loss of motivation at the tails of the pdf, an intermediate step of targeting at the grade level may mitigate the free-rider effect yet retain the desirable cooperative properties. Column 3 in table 7 demonstrates this to be the case. Average teacher effort and student achievement increase. The differences in variance in expected bonus across the three regimes show that targeting at the grade level lowers the number of teachers pushed out to the tails of the probability distribution compared to classroom-level targeting. At the same time, the reduction in the number of teachers to approximately a

[51] The mass at 30 days for classroom regime is an imposed upper limit. Removing the upper limit does not qualitatively change the results.

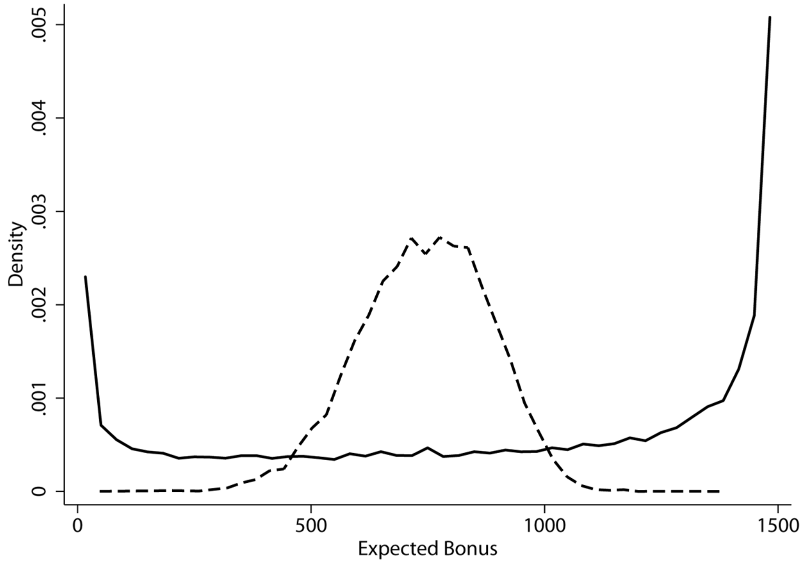

Figure 4.—Histogram of expected bonus under school-level and classroom-level incentives. The dashed line represents school-level incentives. The solid line represents classroom-level incentives.

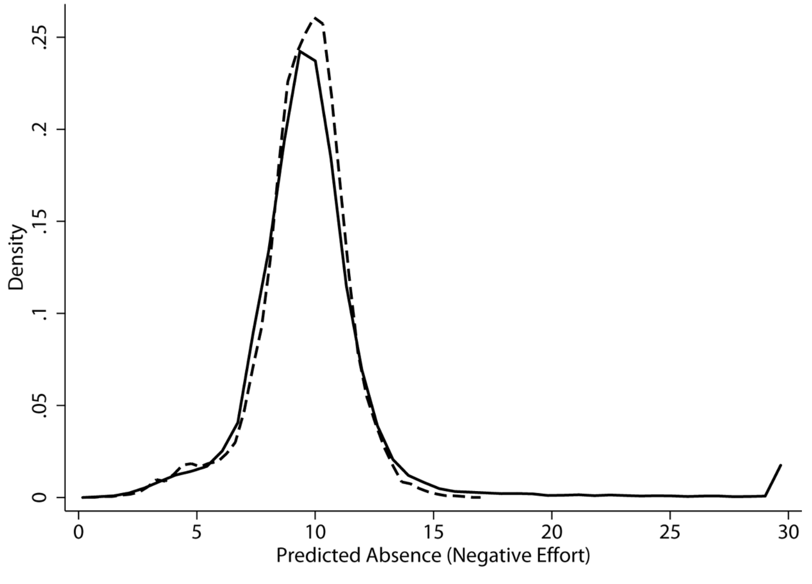

Figure 5.—Histogram of predicted absence under school-level and classroom-level incentives. The dashed line represents school-level incentives. The solid line represents classroom-level incentives.

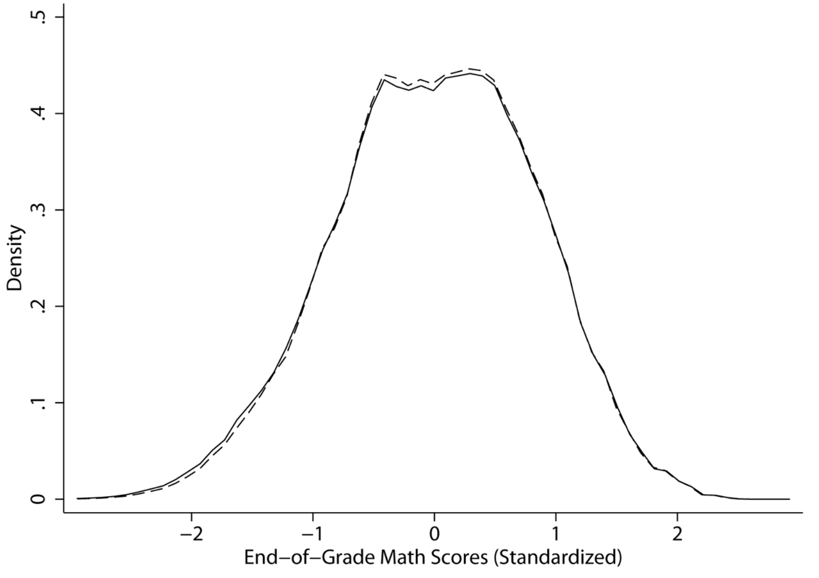

Figure 6.—Histogram of predicted test scores under school-level and classroom-level incentives. The dashed line represents school-level incentives. The solid line represents classroom-level incentives.

third, compared to the school-level targeting, raises the base effort level of all teachers. The combination of these two effects manages to hit a "sweet spot," where it is possible to increase average teacher effort and student achievement.[52] Expected bonus amount for each teacher increases by an amount similar to classroom incentives.

As seen in table 8, the distributional impacts of classroom and grade-level targeting differ for advantaged and disadvantaged students. When the state evaluates bonuses on the basis of classroom averages, the impact on test scores for white students is slightly negative. For students with parents with high levels of education, there is a small positive impact. However, for minority students and students with parents who have a high school education or less, the impact on test scores is large and negative. The last four rows show that roughly one-third of students achieve higher scores under the classroom regime, with students in the more advantaged subgroups having a slight advantage. The benefits of the regime change are concentrated in the middle of the student achievement distribution, while those at the lower end (more heavily concentrated in minority and low–parental education students) and the top end (more heavily concentrated in white and high–parental education students) are deprived. When the regime is changed to grade-level targeting, the gains accrue

[52] This is not meant to imply that grade-level targeting is a first-best solution. There are other alternatives, such as explicitly tying the performance of remedial classes with advanced-track classes to induce academic gains.

TABLE 8
Classroom- and Grade-Level Targeting Distributional Effects

	Classroom (%)	Grade (%)
Mean reading scores change for:		
Students with low parent education	−1.61	+1.02
Minority students	−3.00	+.40
Students with high parent education	+.03	+2.71
White students	−.26	+1.85
Mean math scores change for:		
Students with low parent education	−2.13	+1.34
Minority students	−3.94	+.54
Students with high parent education	+.04	+3.54
White students	−.35	+2.47
% of subgroup with higher scores:*		
Students with low parent education	31.2	97.0
Minority students	31.7	94.3
Students with high parent education	33.7	91.8
White students	35.5	75.1

* Number of students in subgroup whose scores increased as a result of the policy divided by the number of students in the subgroup. Results are identical for reading and math scores.

more evenly across all students. When budgetary concerns are ignored, the policy simulation results would seem to suggest that the state should switch to a grade-level criterion. However, several problems arise in switching to the proposed standard. There could be large, negative general equilibrium effects. Nonrandom sorting of students in grades and classrooms may make it difficult to gauge the teacher's true effectiveness and create tension among teachers (see Rothstein 2010). Teachers may compete to avoid grades least likely to make year-over-year improvements. With experienced teachers having seniority, we may see the students who require the most help ending up with the least experienced teachers. Another possibility is that teachers who are perpetually stuck with underachieving students may seek to transfer or exit the profession altogether. Even if the principal could commit to random sorting of students, some teachers may win or lose out on the bonus by pure chance. In addition, teachers in nontested grades would need to be compensated in some alternative manner. Currently, kindergarten, first-, and second-grade students are not administered an EOG exam. Teachers in charge of these students are paid the bonus according to schoolwide performance. In evaluating a move to an alternative criterion, these negative factors must be considered carefully, along with the positive results from the incentives.

VIII. Conclusion

This study measured the impact of teacher effort on student test scores and examined the effectiveness of accountability legislation using a principal-agent model and the North Carolina administrative education data set. Going beyond reduced-form estimates of the impact of accountability systems on education production, the theoretical and econometric models bridge the gap between academic achievement and the merit pay system by inserting between them the true causal mechanism—teacher effort. Teacher effort is captured by measuring the teacher absence response to changes in incentive pressure, which reduces the noise in raw absence data.

Explicitly modeling the accountability system in this way and estimating a structural econometric model reveals that even a relatively simple and modest accountability program like the North Carolina system can yield complex and unintended consequences. The theoretical model for a pay for performance program in which a bonus is paid to teachers for the school passing a strict threshold for gains in test scores shows that (1) the free-rider problem increases in the number of teachers, (2) the nonlinearity of the probability of bonus receipt leads to higher effort exertions by teachers when the bonus outcome is in doubt, and (3) the free-rider and the incentive effects from the probability of bonus receipt often pull teacher effort in opposite directions. The theoretical model points to the possibility that policy alterations aimed at minimizing the free-rider effect by changing the bonus system from school-level to classroom-level incentives may not be effective in raising teacher effort or student achievement. The parameter estimates from the empirical model confirm that teacher effort (predicted absence) is an important component of student achievement, teachers respond to cash incentives, a free-rider effect exists, and the nonlinearity of the bonus probability is important in determining teacher effort.

I performed two policy simulations to gauge the possible effects of accountability reform. The first experiment changed the criterion from school- to class-level performance, attempting to raise achievement by eliminating the free-rider effect. The second experiment evaluated bonus at the grade level, attempting to balance the free-rider effect and the incentive pressure.

Classroom-level incentives result in lower teacher effort and test score performance compared to school-level incentives. While the free-rider effect is completely gone, separating teachers into individual classes sharply divides many teachers into two extreme groups: those who have almost no chance of qualifying for the bonus and those who are almost assured of the bonus. The rational response of both groups of teachers is to exert less effort. This change in incentive pressure overwhelms the gains in achievement from the elimination of free-rider effects.

Grade-level incentives divide teachers into groups roughly one-third in size. The reduced number of teachers mitigates the impact of free-rider effects, while the grouping still forces the high- and low-ability teachers to exert effort in an attempt to qualify for the bonus. Grade-level incentives yield higher teacher effort and test scores compared to classroom-level or school-level incentives, managing to balance the two forces better than the accountability policies at the two extremes.

The findings in this study hold generally for games in which managers must set all-or-nothing rewards for employees in a cooperative project with noisy output and signals. The North Carolina system is comparable to rewarding the team as a whole on the success of a project. The first simulation moves from team reward to individual contribution. The second simulation adjusts the size of the team to foster collaboration and to mitigate free-riding. Ultimately, the research shows that, as is the case with most well-meaning legislation or incentive systems, there are always unanticipated responses from the targets of the legislation such that the end result may be quite different from what was originally intended.

Appendix A

Proofs

Proof of Existence of a Nash Equilibrium

Since the utility function is concave in effort, the best-response functions can be obtained by solving for first-order conditions. Solving for first-order conditions, J equations, yields

$$B \cdot F'(\cdot)\frac{w_j}{J}\frac{\partial y_j}{\partial e_j} - C'(e_j) = 0.$$

By assumption, a teacher's utility is strictly concave in her own effort. Rewrite the first-order conditions for teacher j as

$$B \cdot F'\left(\frac{w_j \cdot y_j}{J} + Y_{-j}\right)\frac{w_j}{J}\frac{\partial y_j}{\partial e_j} = C'(e_j).$$

Note that $Y_{-j} \in [\underline{Y}, \overline{Y}]$ maps into $e_j \in [\underline{e}, \bar{e}]$ for all j.

Assuming $F''(\cdot) \geq 0$, as we perturb Y from its minimal to maximal value, e_j must increase. That is, the reaction function of e is positively sloped when $F'' \geq 0$. When the reaction function is positively sloped, e is a strategic complement, and it is well established that games in strategic complements have a unique pure-strategy Nash equilibrium. See Milgrom and Shannon (1994) for details.

Assuming $F''(\cdot) < 0$, as we perturb Y from its minimal to maximal value, e_j must decrease. In this case, the reaction function is negatively sloped when $F'' < 0$, and effort is a strategic substitute.

If $Y^L < Y_{-j} < Y^H$ for $i = 1, 2, \ldots, J$, all agents must exert some effort, and there exists a unique pure-strategy Nash equilibrium. See Bramoulle and Kranton (2007) for details. QED.

Proof of a Free-Rider Problem

The condition on x and Y implies an interior solution. Since we assume that all teachers/classes are identical, $y_j = y_{-j}$ for all j. Let x_{J+1}, the additional class/

teacher, also be identical to the other classes. Then Y_{-j} does not change. The first-order condition changes to

$$B \cdot F'\left(\frac{w_j \cdot y_j}{J+1} + Y_{-j}\right)\frac{1}{J+1}\frac{\partial y_j}{\partial e_j} = C'(e_j).$$

This requires e to decrease in order for the first-order condition to be met. Since teachers are identical, all teachers reduce e. The result follows that the incentive system suffers from a free-rider problem. Note here that one cannot take derivatives to test for $\partial e_j/\partial J$ because increases in J are discrete, and an increase in J also implies a new x_{J+1} and e_{J+1}. QED.

Proof of an Individual Criterion Not Always Increasing Effort

Let $J = 2$. Assuming identical class size, in the individual criterion, the first-order conditions are

$$B \cdot F'([e_1 \exp(x_1)])\exp(x_1) - C'(e_1) = 0,$$
$$B \cdot F'([e_2 \exp(x_2)])\exp(x_2) - C'(e_2) = 0.$$

Let $x_1 = x^H$ and $x_2 = x^L$. By assumption 5, $\partial F'/\partial e_i \cong 0$ for $i = 1, 2$. The solution is at a corner, and $e_1^* = e_2^* = 0$, where $\underline{e} = 0$.

In the schoolwide criterion, the first-order conditions are

$$B \cdot F'\left(\left\{\frac{1}{2}[e_1 \exp(x_1) + e_2 \exp(x_2)]\right\}\right)\frac{1}{2}\exp(x_1) - C'(e_1) = 0,$$
$$B \cdot F'\left(\left\{\frac{1}{2}[e_1 \exp(x_1) + e_2 \exp(x_2)]\right\}\right)\frac{1}{2}\exp(x_2) - C'(e_2) = 0,$$

$\partial F'/\partial e_i > 0$. The first-order conditions are jointly satisfied if and only if $e_1^* > 0$ and $e_2^* > 0$. While $e_1^* = e_2^* = 0$ is still a possible solution, the positive effort solution dominates, as $U_i|(e_i = 0) = 0$ for $i = 1, 2$, but $U_i|(e_i \geq 0) \geq 0$ for $i = 1, 2$. QED.

Illustration of the Negative Relationship between Absence and Effort

Assume that teachers get a daily "potential effort" draw from some distribution $G(\delta) \in [0,\bar{\delta}]$ such that teacher i has $\{\delta_{i1}, \delta_{i2}, \ldots, \delta_{id}, \ldots, \delta_{iD}\}$, where D is the total number of days in an academic year. Define μ_i as the mean value of effort for teacher i. There exists some $\underline{\delta}$ such that if $\delta_{id} < \underline{\delta}$, teacher i is absent on day d, $A_{id} = 1$. Therefore, the probability of teacher i being absent on any given day is

$$\Pr(A_{id} = 1) = \Pr(\delta_{id} < \underline{\delta}).$$

Then, the number of absent days for teacher i in year t is $A_{it} = D \cdot \Pr(\delta_{id} < \underline{\delta})$. Assuming that effort is not transferable, potential effort on day d, δ_d, is equivalent to actual effort e_d. For two teachers i and j, $\mu_i \geq \mu_j$ and $E(A_{it}) \leq E(A_{jt})$.

Now assume that effort can be stored but decays at rate λ. For illustration purposes, I focus on the last two days of the academic year before the EOG exam,

$D - 1$ and D. Education production during the two days is $\exp(x)e_{D-1} + \exp(x)e_D$ if the teacher teaches both days. If the teacher opts to take one day off and teach on the last day, education production is $\exp(x)0 + \exp(x)(\lambda e_{D-1} + e_D)$. If $\lambda < 1$, education production declines if effort is stored up. Therefore, if there is any effort decay, education production is maximized by teaching both days.

Now assume that education decays from day to day as students forget material learned. Assume that education decays at rate ξ. Again, focusing on the last two days, education production if the teacher teaches both days is $\exp(x)\xi e_{D-1} + \exp(x)e_D$. If the teacher takes a day off, education production is $\exp(x)\xi 0 + \exp(x)(\lambda e_{D-1} + e_D)$. Teachers have no incentive to store up effort if $\lambda \geq \xi$.

To sum up, if there is no education or effort decay, the assumption $E(A_i) \geq E(A_j)$ if and only if $e_i \leq e_j$ always holds. If effort can be stored, the condition holds if there is any decay in effort from day to day. If effort can be stored and education decays from day to day, the condition holds if the rate of decay of effort is greater than the rate of decay of education. The separability of the daily education production function is not critical to the analysis. In fact, if there is complementarity between days, it is possible to simply define production across both days as equaling $\exp(x)e_{D-1} + \sigma\exp(x)e_D$ if both days are taught and $\exp(x)(e_{D-1} + e_D)$ if only one day of teaching occurs, with $\sigma > 1$. Then the above analysis holds, with σ always pulling the inequality toward teaching both days yielding higher education production.

Decreasing Class Size by Increasing J

Let class achievement of teacher k at time t be redefined as

$$y_{kt} = y(x_{kt}, e_{kt}, sz_{kt}),$$

where sz_{kt} is the class size. For simplicity, class size is the school population divided by J_t, the number of teachers in the schools. For this example, I assume that the school can add "fractional" numbers of teachers, perhaps by adding teaching assistants, substitutes, or team teachers. Then, the probability of bonus receipt is

$$\Pr_t = \Phi\left(\frac{\sum_1^J y(x_{kt}, e_{kt}, sz_{kt})}{J_t} - \overline{Y}_{t-1} > Cr\right).$$

If a principal has control over J, for any given vector of effort by teachers,

$$\begin{aligned}\frac{\partial \Pr}{\partial J_t} &= \Phi'(\cdot)\left[\frac{1}{J}\left(\sum_1^J \frac{\partial y_{kt}}{\partial J_t}\right) + \left(\sum_1^J y_{kt}\right)\left(\frac{-1}{J^2}\right)\right]\\ &= \frac{1}{J}\Phi'(\cdot)\left(\sum_1^J \frac{\partial y_{kt}}{\partial J_t} - \frac{Y_t}{J_t}\right).\end{aligned}$$

Assuming that an increase in class achievement is weakly concave with class size reduction, such that $\partial y/\partial J \geq 0$ and $\partial^2 y/\partial J^2 \leq 0$, for different values of J_t and a given level of effort, $\partial \Pr_t/\partial J_t$ may be greater than, equal to, or less than zero.

While the increase in J increases classroom achievement by decreasing classroom size, it also increases the free-rider effect, which decreases the probability of bonus receipt. It is important to note here that all of these changes to $\Pr_t$ may increase or decrease achievement, and this change may increase or decrease teacher effort, depending on the situation.

If a school is toward the left tail of the probability distribution (the school is essentially out of the bonus) and if $\partial \Pr_t / \partial J_t > 0$, then the principal should hire more teachers if the opportunity is available. The increase in student achievement from reduced class size will move the school toward the middle of the distribution. The response of teachers will then be to exert higher effort, resulting in yet higher achievement. If $\partial \Pr_t / \partial J_t < 0$ for these schools, the free-rider effect overwhelms the positive impact from class size reduction. As the probability decreases even further, teachers have less incentive to exert effort. The optimal response of the principal is to decrease the number of teachers (and increase class size).

On the other hand, if a school is toward the right tail of the probability distribution (bonus is virtually assured) and if $\partial \Pr_t / \partial J_t < 0$, the principal's optimal response is less clear-cut. The free-rider effect is larger than the class size reduction effect, which pulls down test scores (thus decreasing the probability of bonus receipt). The movement to the left will induce more effort on the part of teachers. On the other hand, if $\partial \Pr_t / \partial J_t > 0$ for these schools, again, the principal's best response is unclear. The increase in the number of teachers increases the test score (thus increasing the probability of bonus receipt). However, this increase in probability will elicit less effort from teachers, which will pull achievement down. In both cases, it is unclear if achievement will end up higher or lower than the original score.

Altering Class Size While Leaving J Unchanged

To simplify the analysis, I assume that $J = 2$, the probability CDF is linearized, and class average achievement is linear in effort and class size. Let teacher 1 be the high-ability teacher and let teacher 2 be the low-ability teacher. Teacher j's utility function is

$$U_j = B\Phi(Y) - C(e_j).$$

The effort cost function is simply $C(e_j) = e_j^2$. The schoolwide average achievement Y is defined as

$$\begin{aligned} Y &= f_1 \cdot y_1(f_1, e_1) + f_2 \cdot y_2(f_2, e_2) \\ &= f_1 \cdot y_1(f_1, e_1) + (1 - f_1) \cdot y_2(1 - f_1, e_2), \end{aligned}$$

where f_j is the fraction of the school population assigned to teacher j. Define the linearized CDF such that

$$\Phi = \begin{cases} 0 & \text{if } Y \leq \underline{Y} \\ a + bY & \text{if } \underline{Y} < Y < \overline{Y} \\ 1 & \text{if } Y \geq \overline{Y}. \end{cases}$$

Teacher 1's education production function is

$$y_1 = \alpha_1 + \alpha_2 f_1 + \alpha_3 e_1.$$

Teacher 2's education production function is

$$y_2 = \beta_1 + \beta_2(1 - f_1) + \beta_3 e_2.$$

To impose that teacher 1 has higher ability, I assume that

$$\alpha_1 > \beta_1 > 0,$$
$$0 > \alpha_2 \geq \beta_2,$$
$$\alpha_3 \geq \beta_3 > 0,$$

where α_1 and β_1 represent the base ability of teachers in the absence of additional effort exertion from incentive pressure. The principal attempts to maximize Y by changing the fraction of students taught by teachers 1 and 2 (f_1). Teacher j attempts to maximize U_j by adjusting effort, e_j. Solving for each first-order condition, one can easily see (with the simplifications allowing for closed-form solutions) that

$$e_1^* = B \cdot b(f_1\alpha_3)^2/2,$$
$$e_2^* = B \cdot b[(1 - f_1)\beta_3]^2/2,$$
$$f_1^* = \frac{\beta_1 - \alpha_1 + 2\beta_2 + \beta_3 e_2 - \alpha_3 e_1}{2(\alpha_2 + \beta_2)}.$$

With α_2 and β_2 less than zero, the principal's optimal response is to increase f_1 as teacher 1 becomes more productive (compared to teacher 2). Teacher 1's optimal effort increases as f_1 or α_3 increases, and teacher 2's optimal effort increases as f_1 decreases or β_3 increases.

Because teacher 1 has higher ability, it is reasonable to guess that setting $f_1 = 1$ is optimal. From the first-order condition,

$$f_1^* = 1 \Leftrightarrow \beta_1 - \alpha_1 + 2\beta_2 + \beta_3 e_2 - \alpha_3 e_1 \leq 2(\alpha_2 + \beta_2) < 0.$$

With $f_1 = 1$, we use $e_2^* = 0$, plug in for Y, and simplify to

$$f_1^* = 1 \Leftrightarrow 2 \cdot \beta_1 - Y \leq \alpha_2 < 0.$$

If β_1 is large, then it is possible that $\alpha_2 < 2 \cdot \beta_1 - Y$, and the inequality defined above does not hold. The fact of β_1 (and α_1) being large (relative to α_2, α_3, β_2, and β_3) is a statement about the relative strength of the incentives compared to the base effort of teachers. In this case, if incentives are comparatively weak, principals may want to distribute students more evenly across the two teachers. If the incentive pressure is strong, the principal may be tempted to have teacher 1 teach a higher proportion of (if not all) students.

References

Ahn, T. 2013. "Optimal Matching of Teachers and Schools under Accountability Pressure." Working paper, Dept. Econ., Univ. Kentucky.

Ahn, T., and J. Vigdor. 2012. "How Salient Are Performance Incentives in Education? Evidence from North Carolina." Working paper, Univ. Kentucky and Duke Univ.

———. 2013. "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina." Working paper, Univ. Kentucky and Duke Univ.

Angrist, J., and V. Lavy. 2002. "The Effect of High School Matriculation Awards: Evidence from Randomized Trials." Working Paper no. 9389, NBER, Cambridge, MA.

Barlevy, G., and D. Neal. 2012. "Pay for Percentile." *A.E.R.* 102 (5): 1805–31.

Booher-Jennings, J. 2005. "Below the Bubble: 'Education Triage' and the Texas Accountability System." *American Educ. Res. J.* 42:231–68.

Bradley, S., C. Green, and G. Leeves. 2007. "Worker Absence and Shirking: Evidence from Matched Teacher-School Data." *Labour Econ.* 14 (3): 319–34.

Bramoulle, Y., and R. Kranton. 2007. "Public Goods in Networks." *J. Econ. Theory* 135:478–94.

Carnoy, M., and S. Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educ. Evaluation and Policy Analysis* 24: 305–31.

Chetty, R., J. Friedman, and J. Rockoff. 2011. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." Working Paper no. 17699, NBER, Cambridge, MA.

Clotfelter, C., and H. Ladd. 1996. "Recognizing and Rewarding Success in Public Schools." In *Holding Schools Accountable: Performance-Based Reform in Education*, edited by H. Ladd. Washington, DC: Brookings Inst.

Clotfelter, C., H. Ladd, and J. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *J. Human Resources* 41 (4): 778–820.

———. 2007a. "Are Teacher Absences Worth Worrying about in the U.S.?" Working Paper no. 13648, NBER, Cambridge, MA.

———. 2007b. "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects." Working Paper no. 13617, NBER, Cambridge, MA.

Cullen, J., and R. Reback. 2002. "Tinkering toward Accolade: School Gaming under a Performance Accountability System." Working paper, Univ. Michigan.

Darling-Hammond, L. 2000. "Teacher Quality and Student Achievement: A Review of State Policy Evidence." *Educ. Policy Analysis Archives* 8 (1).

Dee, T. S. 2004. "Teachers, Race, and Student Achievement in a Randomized Experiment." *Rev. Econ. and Statis.* 86 (1): 195–210.

Dee, T. S., and B. Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *J. Policy Analysis and Management* 30:418–46.

Duflo, E., and R. Hanna. 2006. "Monitoring Works: Getting Teachers to Come to School." Discussion Paper no. 5426, Centre Econ. Policy Res., London.

Ehrenberg, R., R. Ehrenberg, D. Rees, and E. Ehrenberg. 1991. "School District Leave Policies, Teacher Absenteeism, and Student Achievement." *J. Human Resources* 26:72–105.

Figlio, D. 2006. "Testing, Crime, and Punishment." *J. Public Econ.* 90:837–51.

Figlio, D., and L. Getzler. 2002. "Accountability, Ability, and Disability: Gaming the System?" Working Paper no. 9307, NBER, Cambridge, MA.

Figlio, D., and L. Kenny. 2007. "Individual Teacher Incentives and Student Performance." *J. Public Econ.* 91 (5–6): 901–14.

Fryer, R. 2011. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." Working Paper no. 16850, NBER, Cambridge, MA.

Glewwe, P., N. Ilias, and M. Kremer. 2010. "Teacher Incentives." *American Econ. J.: Appl. Econ.* 2 (3): 205–27.

Goldhaber, D., and E. Anthony. 2007. "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Rev. Econ. and Statis.* 89 (1): 134–50.

Goldhaber, D., and J. Walch. 2012. "Strategic Pay Reform: A Student Outcomes–Based Evaluation of Denver's ProComp Teacher Pay Initiative." *Econ. Educ. Rev.* 31 (6): 1067–83.

Grissmer, D., and A. Flanagan. 1998. "Exploring Rapid Achievement Gains in North Carolina and Texas." National Education Goals Panel.

Hanushek, E. 2009. "Teacher Deselection." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway, 165–80. Washington, DC: Urban Inst. Press.

Hanushek, E., and M. Raymond. 2005. "Does School Accountability Lead to Student Improvements?" *J. Policy Analysis and Managment* 24:297–327.

Imberman, S., and M. Lovenheim. 2012. "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System." Working Paper no. 18439, NBER, Cambridge, MA.

Jacob, B. 2005. "Accountability, Incentives and Behavior: Evidence from School Reform in Chicago." *J. Public Econ.* 89:761–96.

———. 2007. "Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments." Working Paper no. 12817, NBER, Cambridge, MA.

———. 2010. "The Effect of Employment Protection on Worker Effort: Evidence from Public Schooling." Working Paper no. 15655, NBER, Cambridge, MA.

Jacob, B., and S. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Q.J.E.* 118 (3): 843–77.

Kane, T., and D. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *J. Econ. Perspectives* 16 (4): 91–114.

Ladd, H. 1999. "The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes." *Econ. Educ. Rev.* 18 (1): 1–16.

Lavy, V. 2002. "Evaluating the Effect of Teachers' Performance Incentives on Pupil Achievement." *J.P.E.* 110:1286–1317.

———. 2009. "Performance Pay and Teachers' Effort, Productivity and Grading Ethics." *A.E.R.* 99 (5): 1979–2011.

Macartney, H. 2012. "The Dynamic Effects of Educational Accountability." Working paper, Duke Univ.

Milgrom, P., and C. Shannon. 1994. "Monotone Comparative Statics." *Econometrica* 62:157–80.

Miller, R., R. Murnane, and J. Willett. 2008. "Do Worker Absences Affect Productivity? The Case of Teachers." *Internat. Labour Rev.* 147:71–89.

Muralidharan, K., and V. Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *J.P.E.* 119 (1): 39–77.

Neal, D., and D. Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Rev. Econ. and Statis.* 92 (2): 263–83.

Rivkin, S., E. Hanushek, and J. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73:417–58.

Rockoff, J. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *A.E.R.* 94 (2): 247–52.

Rockoff, J., B. Jacob, T. Kane, and D. Staiger. 2008. "Can You Recognize an Effective Teacher When You Recruit One?" Working Paper no. 14485, NBER, Cambridge, MA.

Rothstein, J. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Q.J.E.* 125 (1): 175–214.

Sanders, W., and J. Rivers. 1996. "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement." Research Progress Report, Value-Added Res. and Assessment Center, Univ. Tennessee.

Sojourner, A., K. West, and E. Mykerezi. 2011. "When Does Teacher Incentive Pay Raise Student Achievement? Evidence from Minnesota's Q-Comp Program." Manuscript, Univ. Minnesota.

Springer, M., et al. 2012. "Team Pay for Performance: Experimental Evidence from the Round Rock Pilot Project on Team Incentives." *Educ. Evaluation and Policy Analysis* 34 (4): 367–90.

Vigdor, J. 2008. "Teacher Salary Bonuses in North Carolina." Working Paper no. 2008-03, Nat. Center Performance Incentives, Vanderbilt Univ.

White, K., and J. Rosenbaum. 2007. "Inside the Black Box of Accountability: How High-Stakes Accountability Alters School Culture and the Classification and Treatment of Students and Teachers." In *No Child Left Behind and the Reduction of the Achievement Gap*, edited by Alan R. Sadovnik, Jennifer A. O'Day, George Bohrnstedt, and Kathryn Borman. New York: Routledge.